A.A. SAMARSKII

MÉTHODES DE RÉSOLUTION DES ÉQUATIONS DE MAILLES

A. SAMARSKI, E. NIKOLAÏEV

MÉTHODES DE RÉSOLUTION DES ÉQUATIONS DE MAILLES

ÉDITIONS MIR · MOSCOU

PRÉFACE

La résolution numérique des équations différentielles de la physique mathématique par la méthode des différences finies s'opère en deux étapes : 1) on procède d'abord à l'approximation discrète de l'équation différentielle sur un maillage (expression du schéma aux différences) ; 2) on résout sur ordinateur les équations aux différences constituant des systèmes d'équations algébriques linéaires d'ordre élevé d'aspect spécial (mauvais conditionnement, structure en bandes de la matrice du système). Il n'est pas toujours rationnel d'appliquer à ces systèmes les méthodes générales de l'algèbre linéaire vu la nécessité de stocker un énorme volume d'information, de même qu'en raison des calculs laborieux que ces méthodes exigent. Pour la résolution des équations aux différences, on développe depuis longtemps des méthodes spéciales qui tiennent plus ou moins compte de la nature spécifique du problème et permettent d'obtenir la solution en un nombre moindre d'opérations, comparé à celui exigé par les méthodes générales de l'algèbre linéaire.

Le présent livre fait suite à celui de A. Samarski et V. Andréev « Méthodes aux différences pour équations elliptiques » étudiant une série de problèmes qui se rapportent à l'approximation discrète, à la construction d'opérateurs de différences et à l'appréciation de la vitesse de convergence des schémas aux différences pour des problèmes aux limites types de la variante elliptique.

Dans ce livre on n'étudie que les méthodes de résolution des équations aux différences. L'ouvrage est de fait divisé en deux parties. La première partie (ch. I-IV) est réservée à l'application des méthodes directes de résolution des équations aux différences, la seconde (ch. V-XV) à la théorie des méthodes itératives de résolution des équations de mailles de forme générale et à leur application aux équations aux différences. Lors de l'utilisation des méthodes directes, un rôle important est attaché à la forme spéciale des équations aux différences. Pour la résolution des équations triponctuelles unidimensionnelles, on s'adresse à différentes variantes de la méthode du balayage (balayage monotone, non monotone, cyclique, en flux, etc.).

Les chapitres III et IV sont consacrés aux méthodes directes économiques appliquées actuellement à la résolution des équations aux différences de Poisson dans un rectangle, associées à des conditions aux limites variées. Citons parmi ces méthodes la méthode de réduction totale et la méthode de séparation des variables utilisant l'algorithme de transformation rapide de Fourier, de même que les méthodes combinées.

L'étude des méthodes itératives s'appuie sur l'assimilation de ces méthodes à des schémas aux différences opératoriels dont le principe a été développé dans les ouvrages de A. Samarski « Introduction à la théorie des schémas aux différences » (1971) et « Théorie des schémas aux différences » (1977). Cette approche autorise d'exposer la théorie des méthodes itératives dans le cadre de la théorie générale de stabilité des schémas aux différences opératoriels sans recourir aux hypothèses sur la structure de la matrice du système (voir de même A. Samarski, A. Gouline « Stabilité des schémas aux différences » (1973)). L'écriture des schémas itératifs sous forme canonique permet non seulement d'isoler les opérateurs responsables de la convergence des itérations mais, également, de confronter les différentes méthodes itératives. Une attention particulière est prêtée à l'étude de la vitesse de convergence des itérations et au choix des paramètres optimaux rendant la vitesse de convergence maximale. La connaissance des estimations de la vitesse de convergence des itérations, comme l'étude de la nature de la stabilité des calculs permettent de procéder dans des situations concrètes à des confrontations des méthodes itératives différentes et de fixer son choix. Bien qu'il soit supposé que le lecteur est initié aux notions sur la théorie des schémas aux différences et aux éléments de l'analyse fonctionnelle, on a jugé utile de donner dans le chapitre V des renseignements élémentaires sur l'appareil mathématique mis en œuvre par la théorie des schémas itératifs et de montrer comment les approximations discrètes des équations elliptiques se réduisent à des équations opératorielles de première espèce $Au = f$, où A sont des opérateurs dans l'espace hilbertien des fonctions de mailles.

Dans les chapitres suivants on étudie le schéma itératif à deux couches muni d'un jeu de paramètres de Tchébychev, assurant la stabilité de calcul de la méthode; le schéma à trois couches; les méthodes itératives du type variationnel (méthodes de la plus grande pente, des moindres résidus, des moindres corrections, des gradients adjoints, etc.); les méthodes itératives pour des équations non autoadjointes et pour le cas d'opérateurs à signes indéterminés et dégénéré; les méthodes des directions alternées; les méthodes « triangulaires » (avec algorithme d'inversion de la matrice triangulaire pour la détermination de la nouvelle itération), telles la méthode de Seidel, la méthode de surrelaxation, etc.; les méthodes itératives de résolution des équations aux différences non linéaires,

la résolution des problèmes aux limites discrets pour des équations elliptiques en coordonnées curvilignes, etc.

Une place particulière est réservée dans ce livre à la méthode universelle triangulaire alternée, mise au point par les auteurs durant les années 1964-1977, dont l'efficience se manifeste de façon particulièrement importante avec la résolution du problème de Dirichlet pour l'équation de Poisson dans un domaine arbitraire et du problème de Dirichlet pour l'équation $\operatorname{div}(k \operatorname{grad} u) = -f(x)$, $x = (x_1, x_2)$ avec des coefficients $k(x)$ fortement variables.

On montre dans le livre comment il faut passer de la théorie générale à des problèmes concrets et l'on y fournit un grand nombre d'algorithmes itératifs pour la résolution des équations aux différences associées à des équations elliptiques et systèmes d'équations. On y donne des estimations du nombre d'itérations et l'on procède à des comparaisons entre des méthodes variées. On montre, en particulier, que pour la résolution d'un problème simple il est plus économique d'utiliser des méthodes directes au lieu de la méthode des directions alternées. Il est opportun de souligner que les problèmes de plus en plus complexes de l'algèbre linéaire, que la pratique nous révèle, rendent urgents aussi bien la mise au point de nouvelles méthodes que l'élargissement du domaine d'application des méthodes anciennes. Le corollaire de ce besoin est la réappréciation des caractéristiques comparées des différentes méthodes.

En écrivant ce livre, les auteurs se sont servis des cours qu'ils ont donnés dans les années 1961-1977 à la faculté de mécanique mathématique et à la faculté de calcul mathématique et de cybernétique de l'Université de Moscou, ainsi que des ouvrages déjà publiés par les auteurs.

Les auteurs profitent de l'occasion pour exprimer leurs remerciements à V. Andréev, I. Friasinov, M. Bakirova, A. Koutchérov, I. Kaporine pour nombre de remarques utiles sur les questions abordées dans le livre.

Les auteurs tiennent de même à exprimer leur gratitude à T. Galichnikova, A. Goloubéva et, tout particulièrement, à V. Martchenko pour l'aide apportée à la préparation du manuscrit.

A. Samarski, E. Nikolaïev

Moscou, décembre 1977

INTRODUCTION

L'utilisation de différentes méthodes numériques (de différences finies, de différences finies du type variationnel, de différences finies et de projections, y compris la méthode des éléments finis) à la résolution des équations différentielles aboutit à un système d'équations algébriques linéaires d'espèce particulière, les équations aux différences. Ce système possède les traits spécifiques suivants: 1) il est d'un ordre élevé, égal au nombre de nœuds possédés par le maillage; 2) le système est mal défini (le rapport de la valeur propre maximale de la matrice associée au système à sa valeur minimale est grand; ainsi, pour l'opérateur de différences de Laplace ce rapport est inversement proportionnel au carré du pas de maillage); 3) la matrice du système est raréfiée, chaque ligne contenant plusieurs éléments différents de zéro dont le nombre est indépendant de celui des nœuds; 4) les éléments non nuls de la matrice sont disposés de façon caractéristique, c'est une matrice bande (matrice de Jacobi).

Dans le calcul approché sur maillage d'équations intégrales et intégro-différentielles on obtient un système d'équations relativement à la fonction donnée sur le maillage (fonction de maille). Il est naturel d'appeler ces équations discrétisées équations de mailles:

$$\sum_{\xi \in \omega} a(x, \xi)\, y(\xi) = f(x), \qquad x \in \omega, \tag{1}$$

où la sommation s'effectue sur tous les nœuds du maillage ω, c'est-à-dire par rapport à un ensemble discret de points. En général, la matrice $(a(x, \xi))$ de l'équation de maille est remplie. Si l'on numérote les nœuds du maillage, l'équation de maille peut s'écrire sous la forme

$$\sum_{j=1}^{N} a_{ij} y_j = f_i, \qquad i = 1, 2, \ldots, N, \tag{2}$$

où i, j sont les numéros des nœuds du maillage, N le nombre total de nœuds. Le cours des raisonnements inverses est évident. L'équa-

tion de maille linéaire est donc un système d'équations algébriques linéaires et, inversement, tout système d'équations algébriques linéaires peut être traité comme une équation de maille linéaire par rapport à une fonction de maille traduite par un maillage dont le nombre de nœuds est égal à l'ordre du système. Remarquons que les méthodes variationnelles (de Ritz, Galerkine et autres) de résolution numérique des équations différentielles conduisent habituellement aux systèmes à matrice remplie.

L'équation aux différences est un cas particulier de l'équation de maille associée à une matrice (a_{ij}) raréfiée. C'est ainsi, par exemple, que (2) est une équation aux différences d'ordre m si sur la ligne au numéro i seul le $m+1$ élément a_{ij} est différent de zéro pour $j = i, \ i+1, \ \ldots, \ i+m$.

Des raisonnements précédents il s'ensuit d'une façon évidente que la résolution des équations de mailles et, en particulier, des équations aux différences relève de l'algèbre linéaire.

* * *

Pour résoudre les problèmes d'algèbre linéaire on utilise un nombre très varié de méthodes numériques à l'amélioration desquelles on travaille sans relâche, certaines étant abandonnées et remplacées par d'autres mieux adaptées. Un grand nombre de méthodes existantes ont ainsi acquis droit de cité et s'appliquent en leur domaine. Aussi pour résoudre un problème concret sur ordinateur doit-on choisir la méthode appropriée parmi l'ensemble de méthodes susceptibles de servir à la résolution du problème donné. La méthode choisie doit, apparemment, posséder les meilleures caractéristiques (ou, comme on a l'habitude de dire, être une méthode optimale), telles que délai minimum d'exécution sur ordinateur (ou minimum d'opérations arithmétiques et logiques nécessaires à la recherche de la solution), stabilité par rapport aux calculs, c'est-à-dire stabilité par rapport aux erreurs dues aux arrondissements, etc.

Il est naturel d'exiger que tout algorithme servant au calcul sur ordinateur soit, en principe, en mesure de fournir la solution du problème donné avec une précision quelconque $\varepsilon > 0$ fixée à l'avance au bout d'un nombre d'opérations $Q(\varepsilon)$. Cette exigence est satisfaite par un ensemble d'algorithmes au sein duquel il s'agit de trouver l'algorithme à minimum $Q(\varepsilon)$ pour tout $\varepsilon > 0$. Un tel algorithme est dit économique. Il est bien entendu que le choix de la méthode « optimale » ou de la « meilleure » méthode s'effectue sur la base de l'ensemble de méthodes connues (et non pas possibles) ; l'expression d'« algorithme optimal » n'a donc qu'un sens limité et conventionnel.

* * *

Le problème de la théorie des méthodes numériques réside dans la recherche de meilleurs algorithmes pour la classe considérée de problèmes, ainsi que dans l'établissement d'une hiérarchie entre les méthodes. La notion même de meilleur algorithme est fonction de l'objectif poursuivi par les calculs.

Le problème du choix de la meilleure méthode peut être posé de deux façons:

a) il s'agit de résoudre un système concret d'équations $Au = f$, $A = (a_{ij})$ étant une matrice;

b) il s'agit de fournir plusieurs variantes de solutions d'un même problème, par exemple, de l'équation $Au = f$ aux seconds membres f variés.

Dans le calcul avec variantes multiples il est possible de diminuer le nombre moyen d'opérations $\overline{Q}(\varepsilon)$ pour une variante en conservant certaines grandeurs et en s'abstenant de les calculer chaque fois de nouveau (par exemple, conserver la matrice inverse).

Il s'ensuit que le choix de l'algorithme doit être guidé par le type de calcul (à une variante ou à plusieurs variantes), les possibilités d'emmagasinage de l'information complémentaire dans la mémoire de l'ordinateur et, partant, par le type de ce dernier, ainsi que par l'ordre du système d'équations. Lorsqu'on apprécie la qualité théorique de l'algorithme de calcul, on se limite habituellement à l'évaluation du nombre d'opérations arithmétiques nécessaires pour obtenir la solution à la précision donnée; dans ce cas le problème des paramètres de l'ordinateur est, en général, négligé.

L'intense développement, ces dernières années, des méthodes numériques de résolution d'équations aux différences approximant les équations différentielles du type elliptique et l'apparition de nouveaux algorithmes économiques ont incité à la révision des conceptions en matière d'applicabilité des méthodes antérieures.

* * *

Le contenu de ce livre est dans une grande mesure conditionné par la nécessité de fournir des méthodes efficaces de résolution d'équations aux différences répondant aux problèmes aux limites pour équations elliptiques de deuxième ordre. Les problèmes aux limites au sens des différences finies peuvent être classés d'après les critères suivants:

1) la forme de l'opérateur différentiel L dans l'équation

$$Lu = f(x), \quad x = (x_1, x_2, \ldots, x_p) \in G; \tag{3}$$

2) la forme du domaine G dans lequel est recherchée la solution;

3) le type des conditions aux limites à la frontière Γ du domaine G;

4) le maillage $\bar{\omega}$ du domaine $\bar{G}=G+\Gamma$ et le schéma aux différences

$$\Lambda y = -\varphi(x), \qquad x \in \omega, \tag{4}$$

c'est-à-dire l'aspect de l'opérateur de différences finies Λ.

En qualité d'exemples d'opérateur elliptique de deuxième ordre on peut citer

$$Lu = \Delta u = \sum_{\alpha=1}^{p} \frac{\partial^2 u}{\partial x_\alpha^2} \text{— opérateur de Laplace,} \tag{5}$$

$$Lu = \sum_{\alpha,\ \beta=1}^{p} \frac{\partial}{\partial x_\alpha}\left(k_{\alpha\beta}(x)\frac{\partial u}{\partial x_\beta}\right) - q(x)\,u, \tag{6}$$

les coefficients $k_{\alpha\beta}(x)$ vérifient en chaque point $x=(x_1, x_2, \ldots, x_p)$ la condition de forte ellipticité

$$c_1\sum_{\alpha=1}^{p}\xi_\alpha^2 \leqslant \sum_{\alpha,\ \beta=1}^{p} k_{\alpha\beta}(x)\,\xi_\alpha\xi_\beta \leqslant c_2\sum_{\alpha=1}^{p}\xi_\alpha^2, \qquad c_1,\ c_2=\text{const}>0, \tag{7}$$

où $\xi=(\xi_1, \ldots, \xi_p)$ est un vecteur quelconque. Si $u(x)\approx \approx (u^1(x), u^2(x), \ldots, u^m(x))$ est le vecteur-fonction, alors (3) constitue un système d'équations et

$$(Lu)^i = \sum_{j=1}^{m}\sum_{\alpha,\ \beta=1}^{p} \frac{\partial}{\partial x_\alpha}\left(k_{\alpha\beta}^{ij}\frac{\partial u^j}{\partial x_\beta}\right), \quad i=1,\ 2,\ \ldots,\ m,$$

quant à la condition de forte ellipticité, elle prend la forme

$$c_1\sum_{i=1}^{m}\sum_{\alpha=1}^{p}(\xi_\alpha^i)^2 \leqslant \sum_{i,\ j=1}^{m}\ \sum_{\alpha,\ \beta=1}^{p} k_{\alpha\beta}^{ij}(x)\,\xi_\alpha^i\xi_\beta^j \leqslant c_2\sum_{i=1}^{p}\sum_{\alpha=1}^{m}(\xi_\alpha^i)^2,$$

$$c_1,\ c_2=\text{const}>0.$$

* * *

La forme du domaine exerce une forte influence sur les propriétés de la matrice des équations aux différences. On dégagera les domaines pour lesquels l'équation $Lu=0$ aux conditions aux limites homogènes autorise la séparation des variables. C'est ainsi que pour l'équation de Laplace en coordonnées cartésiennes (x_1, x_2) $Lu=\Delta u= = \frac{d^2u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}$, la méthode de séparation des variables est applicable au cas où G est un rectangle. Une propriété analogue possède également le schéma aux différences sur un maillage rectangle, par exemple, le schéma « croix » ; le maillage peut dans ce cas présenter des irrégularités dans chaque direction.

Pour la confrontation des différentes méthodes numériques de résolution de systèmes d'équations algébriques on utilisera en qualité d'*étalon* ou de *modèle* de problème le problème de différences aux limites suivant:

équation de Poisson, domaine — un carré, conditions aux limites de première espèce, maillage carré de pas $h_1 = h$ et $h_2 = h$ suivant x_1 et x_2, opérateur de différences Λ à cinq points.

Le deuxième groupe de problèmes de différences aux limites répond aux données suivantes: L — opérateur aux coefficients variables de la forme (6): a) sans dérivées mixtes, b) avec dérivées mixtes, domaine $G = \{0 \leqslant x_\alpha \leqslant l_\alpha, \quad \alpha = 1, 2\}$ — rectangle (parallélépipède pour $p \geqslant 3$).

Le troisième groupe de problèmes a un domaine de forme compliquée, quant à L, c'est soit l'opérateur de Laplace, soit un opérateur de forme commune; la complicité du problème est en premier lieu fonction de la forme du domaine, du choix du maillage et de l'opérateur de différences au voisinage de la frontière.

Pour le deuxième et le troisième groupe de problèmes on choisit habituellement l'opérateur de différences de manière à sauvegarder les principales propriétés (autoconjugaison, signes définis, etc.) du problème initial et à satisfaire à l'exigence de l'approximation avec un ordre déterminé relativement au pas du maillage.

* * *

Pour la résolution de problèmes de différences elliptiques on recourt à des méthodes directes et itératives.

Les méthodes directes sont applicables dans des cas multidimensionnels, essentiellement pour les problèmes du premier groupe (L — opérateur de Laplace, G — rectangle pour $p = 2$ et parallélépipède pour $p \geqslant 3$, Λ — schéma aux différences à cinq ou neuf points pour $p = 2$). Au cas de problèmes unidimensionnels, quand l'équation aux différences est de second ordre (la matrice est tridiagonale), les coefficients de l'équation pouvant varier, on peut utiliser la méthode du balayage qui est une variante de la méthode de Gauss (voir ch. II). Il existe une série de variantes de la méthode du balayage: balayage monotone, balayage non monotone, balayage en flux, balayage cyclique, etc. (voir ch. II). Pour les problèmes à deux dimensions du premier groupe (voir plus haut), la méthode efficace est celle de réduction totale (ch. III), la méthode de séparation des variables avec la transformation rapide de Fourier, de même que cele combinant la méthode de réduction incomplète avec la transformation rapide de Fourier (ch. IV). Dans tous les cas, suivant une des directions on utilise la méthode de balayage pour la résolution de l'équation aux différences de deuxième ordre.

Les méthodes directes indiquées au cas du problème de diffé-

rences de Dirichlet pour l'équation de Poisson dans un rectangle $(0 \leqslant x_\alpha \leqslant l_\alpha,\ \alpha = 1, 2)$ sur un maillage $\bar{\omega} = \{(i_1h_1, i_2h_2),\ i_\alpha = 0, 1, \ldots, N_\alpha,\quad h_\alpha = l_\alpha/N_\alpha,\quad \alpha = 1, 2\}$ exigent $Q = O(N_1N_2 \log_2 N_2)$ opérations arithmétiques, où $N_2 = 2^n$, $n > 0$ est un nombre entier.

Les méthodes directes servent pour des problèmes d'une classe très spéciale.

* * *

Les problèmes de différences elliptiques au cas où les opérateurs L sont de l'espèce commune ou bien les domaines sont de forme compliquée sont résolus essentiellement à l'aide des méthodes itératives.

Les équations de mailles peuvent être assimilées à des équations opératorielles de première espèce

$$Au = f \tag{8}$$

aux opérateurs définis sur les espaces H des fonctions de mailles. Dans l'espace H on introduit le produit scalaire (.) et les normes énergie $\| u \|_D = \sqrt{(Du, u)}$, $D = D^* > 0$, $D: H \rightarrow H$, où D est un certain opérateur linéaire dans H.

Les méthodes itératives de résolution de l'équation opératorielle $Au = f$ peuvent être traitées comme opératorielles au sens des différences finies (relativement au temps fictif ou au numéro-indice d'itération) de l'équation aux opérateurs dans l'espace hilbertien H. Si la suivante itération y_{k+1} se calcule après m itérations précédentes $y_k, y_{k-1}, \ldots, y_{k-m+1}$, la méthode itérative (schéma) est appelée méthode à $m + 1$ couches (à m pas). Il s'ensuit l'analogie des schémas itératifs avec les schémas aux différences pour problèmes non stationnaires. Aussi la théorie des méthodes itératives constitue-t-elle de même une branche spéciale de la théorie générale de la stabilité des schémas aux différences opératoriels. On se bornera à l'étude de schémas à deux couches et dans une moindre mesure de ceux à trois couches. Le passage aux schémas multicouches n'implique aucun avantage (comme d'ailleurs il s'ensuit de la théorie générale de la stabilité, voir [10]).

Un rôle important revient à l'écriture des méthodes itératives en une forme unique (canonique) permettant de séparer l'opérateur (le stabilisateur) régissant la stabilité et la convergence des itérations et de comparer les différentes méthodes itératives sur la base de critères communs.

Toute méthode itérative à deux couches (à un pas) s'écrit sous la forme canonique suivante:

$$B\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f,\qquad k = 0, 1, \ldots,\qquad y_0 \in H, \tag{9}$$

où $B: H \to H$ est l'opérateur linéaire muni de B^{-1} inverse, τ_1, τ_2, . . . étant les paramètres d'itérations, k le numéro d'itération, y_k l'approximation itérative de numéro k. Dans le cas général $B = B_{k+1}$ dépend de k. Dans la théorie générale on admet que B ne dépend pas de k.

Les paramètres $\{\tau_k\}$ et l'opérateur B sont quelconques et ils doivent être choisis en partant de la condition du minimum d'itérations n, pour lequel la solution y_n de l'équation (9) approche dans H_D la solution précise u de l'équation $Au = f$ avec une précision relative $\varepsilon > 0$:

$$\| y_n - u \|_D \leqslant \varepsilon \| y_0 - u \|_D. \tag{10}$$

Pour la théorie générale des méthodes itératives exposée dans ce livre il n'est pas nécessaire de faire des hypothèses sur la structure de l'opérateur A (de la matrice (a_{ij})). On n'utilise que les propriétés de forme générale

$$A = A^* > 0, \quad B = B^* > 0, \quad \gamma_1 B \leqslant A \leqslant \gamma_2 B, \quad \gamma_1 > 0. \tag{11}$$

Les inégalités opératorielles signifient que sont définies les constantes γ_1, γ_2 de l'équivalence énergétique des opérateurs A et B ou les bornes du spectre de l'opérateur A dans l'espace H_B (γ_1 et γ_2 sont les valeurs propres minimale et maximale du problème généralisé relativement aux valeurs propres: $Av = \lambda Bv$).

* * *

La solution $\tau_1, \tau_2, \ldots, \tau_n$ du problème mentionné plus haut du $\min\limits_{\tau_1, \tau_2, \ldots, \tau_n} n_0(\varepsilon)$ pour des γ_1, γ_2 donnés et un B fixé au cas où $D = AB^{-1}A$ s'exprime au moyen des zéros du polynôme de Tchébychev d'ordre n (méthode itérative de Tchébychev). Pour ces valeurs optimales de $\tau_1, \tau_2, \ldots, \tau_n$ et $\varepsilon > 0$ arbitrairement défini pour n itérations calculées suivant le schéma (9), l'estimation $n \geqslant \dfrac{\ln(2/\varepsilon)}{\ln((1+\sqrt{\xi})/(1-\sqrt{\xi}))}$ ou $n \geqslant n_0(\varepsilon) = \dfrac{\ln(2/\varepsilon)}{2\sqrt{\xi}}$, $\xi = \gamma_1/\gamma_2$ est vraie, tandis que l'inégalité

$$\| Ay_n - f \|_{B^{-1}} \leqslant \varepsilon \| Ay_0 - f \|_{B^{-1}} \quad .$$

est satisfaite.

La stabilité de calcul de la méthode de Tchébychev a lieu au cas d'un mode particulier de numération (de mise en ordre) des zéros du polynôme de Tchébychev et des paramètres $\tau_1^*, \tau_2^*, \ldots, \tau_n^*$; ce mode est décrit au ch. VI.

Pour $B = E$ (E est l'opérateur unité) la méthode (9) est dite explicite, tandis que pour $B \neq E$, elle est implicite. Si le paramètre τ_k est choisi constant, $\tau_k = \tau_0 = 2/(\gamma_1 + \gamma_2)$, $k = 1, 2, \ldots, n$,

on obtient alors un schéma implicite d'une simple itération, pour lequel $n \geqslant n_0(\varepsilon) = \ln\left(\frac{1}{\varepsilon}\right) \Big/ (2\xi)$.

L'opérateur B (stabilisateur) est choisi en fonction des raisons économiques, c'est-à-dire du minimum de travail de calcul au cours de la résolution de l'équation $Bv = F$, le second membre F étant donné, et, comme il a été déjà mentionné, en fonction de la condition du minimum d'opérations d'itérations $n_0(\varepsilon)$.

Supposons qu'on est en mesure de résoudre économiquement le problème $Rv = f$ en effectuant $Q_R(\varepsilon)$ opérations, où

$$R\colon H \to H, \quad R = R^* > 0, \quad c_1 R \leqslant A \leqslant c_2 R, \quad c_1 > 0. \tag{12}$$

On peut alors poser $B = R$ et chercher la solution du problème $Au = f$ suivant le schéma (9) avec les paramètres $\{\tau_k^*\}$ pour $\gamma_1 = c_1$, $\gamma_2 = c_2$ en $Q_A(\varepsilon) \approx \frac{1}{2}\sqrt{c_2/c_1}\ln(2/\varepsilon)\, Q_R(\varepsilon)$ opérations.

Si, par exemple, L est l'opérateur de forme générale, G le rectangle, on peut prendre en guise de R l'opérateur aux différences de Laplace pentaponctuel et résoudre l'équation $Rv = f$ par la méthode directe.

Il peut s'avérer commode de résoudre l'équation $Rv = f$ non pas de façon directe, mais par la méthode itérative; dans ce cas $B \neq R$ et ne s'exprime pas sous forme explicite mais se retrouve par voie itérative.

* * *

Les méthodes connues de Zeidel et de surrelaxation sont des méthodes implicites et correspondent aux matrices triangulaires (opérateurs) B. La convergence de ces méthodes se démontre sur la base de la théorie générale de schémas aux différences (voir A. A. Samarski « Theoria raznostnykh skhem » (« Théorie des schémas aux différences ») M., 1977 ou A. A. Samarski, A. V. Gouline « Ustoïtchivost raznostnykh skhem » (« Stabilité des schémas aux différences ») M., 1973). Toutefois, en ce qui concerne les méthodes de Zeidel et de surrelaxation l'opérateur B n'est pas autoadjoint, aussi ne peut-on profiter de la méthode de Tchébychev (9) au choix optimal des paramètres d'itérations $\tau_1^*, \tau_2^*, \ldots, \tau_n^*$, ce qui permettrait d'augmenter la célérité de convergence des itérations. L'opérateur B peut être rendu autoadjoint en posant qu'il est égal au produit d'opérateurs mutuellement adjoints

$$B = (E + \omega R_1)(E + \omega R_2), \quad R_2^* = R_1, \tag{13}$$

où $\omega > 0$ est un paramètre. Pour R_1 et R_2 on peut prendre des opérateurs possédant des matrices triangulaires (R_1 inférieure et R_2 supérieure), de sorte que $R_1 + R_2 = R\colon H \to H$, $R^* = R > 0$. En particulier, on peut poser

$$R_1 + R_2 = A, \quad R_2^* = R_1. \tag{14}$$

Les hypothèses typiques sont

$$R \geqslant \delta E, \quad R_1R_2 \leqslant \frac{\Delta}{4} A, \quad \delta > 0, \ \Delta > 0. \qquad (15)$$

Ensuite, en choisissant $\omega = 2/\sqrt{\delta\Delta}$ à partir de la condition min $n_0(\varepsilon)$, on obtient les paramètres γ_1, γ_2 et l'on calcule les paramètres $\{\tau_k^*\}$. La détermination de y_{k+1} au moyen de y_k et f se réduit à la résolution successive de deux systèmes d'équations avec matrices triangulaires inférieure et supérieure.

Appelons la méthode itérative (9) construite avec opérateur B factorisé sous forme (13) méthode triangulaire alternée (MTA). La méthode MTA est évidemment une méthode universelle, vu que la représentation de A sous forme de somme $R_1 + R_2 = A$, $R_2^* = R_1$ est toujours possible. Au cas d'un problème de différences elliptique la construction de R_1 et R_2 s'avère aisée. Ainsi, par exemple, $R_1y \rightarrow \sum\limits_{\alpha=1}^{p} \frac{y_{\bar{x}_\alpha}}{h_\alpha}$, $R_2y \rightarrow -\sum\limits_{\alpha=1}^{p} \frac{y_{x_\alpha}}{h_\alpha}$, au cas où Ay est un opérateur de différences de Laplace à $2p + 1$ points, $Ay \rightarrow -\sum\limits_{\alpha=1}^{p} y_{\bar{x}_\alpha x_\alpha}$, h_α étant le pas du maillage en direction de Ox_α. Cette méthode se caractérise par une rapide convergence. En prenant l'ensemble des paramètres de Tchébychev $\{\tau_k^*\}$ et compte tenu de (14), (15), on a alors pour le nombre d'itérations de la MTA l'estimation

$$n_0(\varepsilon) \geqslant \frac{1}{2\sqrt{2}\sqrt[4]{\eta}} \ln \frac{2}{\varepsilon}, \qquad \eta = \frac{\delta}{\Delta}. \qquad (16)$$

En particulier, pour le problème modèle on a $n \geqslant n_0(\varepsilon) = 0{,}3 \ln \frac{2}{\varepsilon} \Big/ \sqrt{h}$.

Au cas d'un domaine quelconque et des équations à coefficients variables, il est logique d'utiliser la méthode triangulaire alternée modifiée (MTAM) en posant

$$B = (\mathscr{D} + \omega R_1)\, \mathscr{D}^{-1} (\mathscr{D} + \omega R_2), \quad R_2^* = R_1, \quad \mathscr{D} = \mathscr{D}^* > 0, \qquad (17)$$

où $\mathscr{D}$ est un opérateur arbitraire. Si à la place de (15) sont satisfaites les inégalités

$$R \geqslant \delta \mathscr{D}, \quad R_1 \mathscr{D}^{-1} R_2 \leqslant \frac{\Delta}{4} \mathscr{D}, \quad \delta > 0, \quad \Delta > 0, \qquad (18)$$

l'estimation (16) reste valable.

On définit ici δ et Δ en choisissant l'opérateur $\mathscr{D}$ et le paramètre ω de manière que le rapport $\xi = \gamma_1/\gamma_2$ soit maximal. En pratique en guise de matrice $\mathscr{D}$ on peut choisir une matrice diagonale.

Indiquons deux exemples d'application efficace de la MTAM.

1) Problème de Dirichlet pour l'équation de Poisson dans un domaine à deux dimensions de forme compliquée; le maillage prin-

cipal dans le plan (x_1, x_2) est régulier. de pas h, le schéma étant à cinq points. Avec un choix correspondant de $\mathcal{D}$, la MTAM n'exige pas plus de 4 à 5 % d'itérations supplémentaires par rapport au même problème dans un carré de côté égal au diamètre du domaine.

2) Dans des équations elliptiques aux coefficients variant fortement (le rapport c_2/c_1 est grand) la MTAM associée à un $\mathcal{D}$ choisi convenablement permet d'affaiblir la dépendance de c_2/c_1.

En pratique on recourt, à côté de méthodes (9) à un pas (à deux couches), à des schémas itératifs à deux pas (à trois couches). Au cas de paramètres itératifs optimaux, le nombre d'itérations devient comparable à celui du schéma de Tchébychev à paramètres $\{\tau_k^*\}$ quand $\xi \to 0$, toutefois, ces méthodes sont plus sensibles relativement aux erreurs de détermination de γ_1 et γ_2. Si les conditions (11) sont remplies, il vaut mieux se servir du schéma de Tchébychev (9) à paramètres $\{\tau_k^*\}$.

* * *

Dans la résolution des problèmes elliptiques un rôle très important revient à la méthode itérative des directions alternées (MDA) développée, à partir de 1955, par de nombreux auteurs. Elle ne s'est toutefois avérée économique que pour une classe très réduite de problèmes du premier groupe, lorsque les conditions $A = A_1 + A_2$, $A_\alpha = A_\alpha^* \geqslant 0$, $\alpha = 1, 2$, $A = A^* > 0$, $A_1A_2 = A_2A_1$ sont remplies. Si A_1 et A_2 sont permutables, on peut choisir pour la MDA des paramètres itératifs optimaux. Pour un problème modèle muni de ces paramètres le nombre d'itérations $n_0(\varepsilon) = O\left(\ln \frac{1}{h} \ln \frac{1}{\varepsilon}\right)$, quant à celui d'opérations, il s'élève à $Q(\varepsilon) = O\left(\frac{1}{h^2} \ln \frac{1}{h} \ln \frac{1}{\varepsilon}\right)$, tandis que pour les méthodes directes $Q = O\left(\frac{1}{h^2} \ln \frac{1}{h}\right)$. Dans ce cas les méthodes directes sont plus économiques que la MDA. Si A_1 et A_2 ne sont pas permutables, la MDA exige $O\left(\frac{1}{h} \ln \frac{1}{\varepsilon}\right)$ itérations, tandis qu'avec la MTA on peut se limiter à $O\left(\frac{1}{\sqrt{h}} \ln \frac{1}{\varepsilon}\right)$ itérations. Au cas de problèmes tridimensionnels, quand $A = A_1 + A_2 + A_3$, même dans l'hypothèse d'une permutabilité deux à deux de A_1, A_2, A_3, la MDA exige plus d'opérations que la MTA. Aussi la MDA a-t-elle perdu pour une grande part son importance.

* * *

Si l'opérateur $A > 0$ n'est pas autoadjoint, il est impossible de construire pour $A = A^* > 0$ le processus itératif de même vitesse de convergence que la méthode de Tchébychev au moyen du schéma

(9) muni de l'assortiment de paramètres et de l'opérateur autoadjoint $B = B^* > 0$. Toutes les méthodes connues ont une vitesse de convergence inférieure. On étudie ici la méthode itérative simple (ch. VI) en définissant à priori deux sortes d'information:

a) on définit les paramètres γ_1, γ_2 figurant dans les conditions (pour simplifier, posons $D = B = E$)

$$\gamma_1 (x, x) \leqslant (Ax, x), \quad (Ax, Ax) \leqslant \gamma_2 (Ax, x), \quad \gamma_1 > 0, \quad \gamma_2 > 0; \quad (19)$$

b) trois paramètres sont définis γ_1, γ_2, γ_3, où γ_1 et γ_2 (si $D = B = E$) sont les bornes de la partie symétrique de l'opérateur A:

$$\gamma_1 E \leqslant A \leqslant \gamma_2 E, \quad \| A_1 \| \leqslant \gamma_3, \quad \gamma_1 > 0, \quad \gamma_2 \geqslant 0, \quad (20)$$

où $A_1 = 0,5 (A - A^*)$ est la partie symétrique gauche de A.

En choisissant τ à partir de la condition du minimum de la norme de l'opérateur de transition ou de l'opérateur résolvant, on aboutit dans tous les cas à l'augmentation du nombre d'itérations comparé au cas de $A = A^*$.

* * *

Toute méthode itérative à deux couches construite sur la base du schéma (9) est caractérisée par les opérateurs B et A, un espace énergie H_D pour lequel on démontre la convergence de la méthode et un assortiment de paramètres. Si l'opérateur B est fixé, le problème principal se ramène alors à la recherche de $\{\tau_k\}$.

Avec le choix des paramètres $\{\tau_k\}$ on utilise l'information à priori sur les opérateurs du schéma. L'aspect de l'information est fonction des propriétés des opérateurs A, B et D. C'est ainsi que dans le cas du schéma de Tchébychev pour $D = AB^{-1}A$, quand A et B sont des opérateurs autoadjoints, on admet que les constantes γ_1, γ_2 de (11) sont données. Dans le cas général, quand $DB^{-1}A$ est autoadjoint dans H, au lieu de (11) il suffit d'exiger que $\gamma_1 D \leqslant DB^{-1}A \leqslant \gamma_2 D$, $\gamma_1 > 0$. Dans le cas de non-autoconjugaison, quand $A \neq A^*$, tandis que $B = B^* > 0$, on utilise soit deux nombres γ_1, γ_2, soit trois nombres γ_1, γ_2 (entrant dans (19)) et γ_3, constante entrant dans la partie symétrique gauche de l'opérateur A. Dans nombre de cas la recherche des constantes γ_1, γ_2 et γ_3 avec une suffisante précision peut devenir un problème autonome assez compliqué impliquant la résolution d'algorithmes spéciaux. Si l'information à priori peut être obtenue au prix de quelques calculs ou si des calculs multiples sont exigés pour la résolution de l'équation $Au = f$ à seconds membres variés, il est logique de rechercher une fois pour toutes les nombres exigés γ_1, γ_2, γ_3 et d'utiliser ensuite la méthode de Tchébychev ou la MTA. S'il s'agit de ne résoudre que le problème $Au = f$ et si l'approximation initiale est bonne, tandis que le calcul des constantes γ_1, γ_2 s'avère laborieux, il faut recourir aux méthodes itératives du type variationnel.

Les méthodes itératives du type variationnel n'impliquent pas la connaissance de γ_1, γ_2 lorsqu'on calcule les paramètres $\{\tau_k\}$. Ces méthodes n'utilisent que l'information de forme générale

$$A = A^* > 0, \quad (DB^{-1}A)^* = DB^{-1}A. \tag{21}$$

Pour déterminer y_{k+1}, on utilise le même schéma (9) en ne modifiant que la formule de τ_{k+1}. Le paramètre τ_{k+1} s'obtient à partir de la condition du minimum dans H_D de la norme d'erreur $z_{k+1} = y_{k+1} - u$, c'est-à-dire du minimum de la fonctionnelle $I[y] = (D(y-u), y-u)$. Le paramètre τ_{k+1} se calcule sur la base de y_k. En choisissant $D = A$, on obtient la méthode de la descente la plus rapide, tandis que pour $D = A^*A$ on a la méthode d'écarts minimums, etc. Ces méthodes sont de même rapidité de convergence que la méthode itérative simple (avec constantes γ_1, γ_2 précises). La rapidité de convergence peut être élevée si l'on s'abstient de minimiser localement (par pas) $\| z_{k+1} \|_D$ et l'on choisit les paramètres τ_k à partir de la condition de minimisation de la norme d'erreur $\| z_n \|_D$ pour tous les n pas, c'est-à-dire en passant de y_0, à y_n. Ce procédé conduit aux schémas itératifs de directions conjuguées (gradients, écarts, corrections ou erreurs conjugués) à deux paramètres (associés à chaque k) et à trois couches, dont la rapidité de convergence est la même que celle de la méthode de Tchébychev à paramètres $\{\tau_k^*\}$ calculés d'après les valeurs exactes de γ_1, γ_2. Si $A = A^* > 0$, on est en mesure de reproduire le processus d'accélération ($\approx$ de 1,5 à 2 fois) de la convergence des méthodes du gradient à deux couches.

* * *

En théorie générale des méthodes itératives il n'est pas exigé de connaître la structure concrète des opérateurs du problème; on n'utilise qu'un minimum d'information générale de nature fonctionnelle sur les opérateurs, par exemple, les conditions (11). Le choix de l'opérateur B du schéma (9) doit se plier aux exigences: 1) d'assurer à la méthode (9) la convergence la plus rapide, 2) de veiller à l'inversion économique de B. Dans la construction de B on peut partir de l'opérateur $R = R^* > 0$ (régularisateur) à énergie équivalente $A = A^* > 0$, $B = B^* > 0$:

$$c_1 R \leqslant A \leqslant c_2 R, \quad c_1 > 0, \quad \mathring{\gamma}_1 B \leqslant R \leqslant \mathring{\gamma}_2 B, \quad \mathring{\gamma}_1 > 0. \tag{22}$$

De sorte que $\gamma_1 = c_1 \mathring{\gamma}_1$, $\gamma_2 = c_2 \mathring{\gamma}_2$. Pour des A différents on peut choisir un même régularisateur R. Le plus souvent on se trouve en présence d'un opérateur B factorisé, par exemple,

$$B = (E + \omega R_1)(E + \omega R_2), \quad R_1 + R_2 = R, \tag{23}$$

où

$$R_1^* = R_2 > 0 \qquad \text{pour MTA}, \tag{24}$$

$$R_1^* = R_1 > 0, \quad R_2^* = R_2 > 0, \quad R_1 R_2 = R_2 R_1 \quad \text{pour MDA}. \tag{25}$$

Pour appliquer la théorie, il faut trouver $\mathring{\gamma}_1$ et $\mathring{\gamma}_2$; le paramètre $\omega > 0$ s'obtient à partir de la condition du min $(\mathring{\gamma}_1(\omega)/\mathring{\gamma}_2(\omega))$. Si l'équation $Rw = F$ se prête à une résolution économique par la méthode directe, on pose alors $B = R$ (par exemple, au cas où $(-R)$ est un opérateur de différence de Laplace, le domaine un rectangle). L'opérateur B peut ne pas s'écrire de façon explicite, mais peut intervenir dans la résolution itérative de l'équation $Rw = r_k$, $r_k = Ay_k - f$ (méthode de deux étapes).

* * *

Pour les équations aux opérateurs A à signes indéterminés, dégénérés et complexes, on peut utiliser les mêmes schémas (9). Cependant le choix des paramètres optimaux se complique, tandis que la rapidité de convergence diminue. Un « traitement » préalable du problème initial est nécessaire avant l'application de la théorie générale à ces cas particuliers. Il s'avère possible de construire des modifications de la méthode de Tchébychev, comme des méthodes du type variationnel.

Si A est un opérateur linéaire dégénéré, c'est-à-dire que l'équation homogène $Au = 0$ possède une solution non triviale, le problème (9) pour $B = E$ et τ_k quelconques a toujours une solution. Soit $H^{(0)}$ le sous-espace associé à la valeur propre zéro de l'opérateur A, $H^{(1)}$ le complément orthogonal de $H^{(0)}$ jusqu'à H. Tout vecteur $y \in H^{(0)}$ vérifie l'équation $Ay = 0$. Si $f \in H^{(1)}$ et $y_0 \in H^{(1)}$, alors toutes les itérations $y_k \in H^{(1)}$. Les conditions

$$\gamma_1 (y, y) \leqslant (Ay, y) \leqslant \gamma_2 (y, y), \quad y \in H^{(1)}, \quad \gamma_1 > 0$$

étant remplies, on peut utiliser le schéma explicite (9) avec les paramètres $\{\tau_k^*\}$ de Tchébychev obtenus sur la base de γ_1, γ_2. Dans ce cas y_k converge vers la solution normale possédant la norme minimale.

Si $f = f^{(0)} + f^{(1)}$ et $f^{(0)} \neq 0$, en entendra par solution normale généralisée de l'équation $Au = f$ la solution de l'équation $Au^{(1)} = f^{(1)}$, $u^{(1)} \in H^{(1)}$, possédant une norme minimale. L'estimation

$$\| y_n - u^{(1)} \| \leqslant \tilde{q}_n \| y_0 - u^{(1)} \|, \qquad \tilde{q}_n = q_{n-1} (1 + (n-1)) \sqrt{\frac{1 - q_{n-1}^2}{\xi}},$$

$$q_n = \frac{2\rho_1^n}{1 + \rho_1^{2n}}, \qquad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \qquad \xi = \frac{\gamma_1}{\gamma_2}, \quad y_n, \ y_0 \in H^{(1)},$$

est vraie si $\tau_1^*, \tau_2^*, \ldots, \tau_{n-1}^*$ sont les paramètres de Tchébychev, tandis que $\tau_n^* = -\sum_{j=1}^{n-1} \tau_j^*$. La rapidité de convergence diminue par rapport à la convergence dans le cas de A non dégénéré avec les

mêmes γ_1, γ_2. A côté de la méthode de Tchébychev modifiée, mentionnée plus haut, il est possible d'appliquer des méthodes du type variationnel.

La théorie générale permet d'étudier le schéma implicite d'une itération simple pour le cas où H est un espace hilbertien complexe, $A = \tilde{A} + qE$, $\tilde{A}$ étant l'opérateur hermitien, $q = q_1 + iq_2$ un nombre complexe, le paramètre d'itération prenant une valeur optimale. Le passage à la méthode des directions alternées s'effectue également sans peine.

* * *

Les résultats de la théorie générale peuvent être appliqués sans difficultés à la résolution des équations aux différences approximant les problèmes aux limites pour des équations du type elliptique. Il est dans ce cas facile d'énoncer les règles générales de résolution des problèmes de différences. Soit une équation aux différences $Au = f$, où $A: H \to H$ est l'opérateur de différences défini dans l'espace H des fonctions de mailles associées au maillage ω. D'abord on étudie les propriétés générales de l'opérateur A en établissant, par exemple, sa autoconjugaison et sa positivité, $A = A^* > 0$, ensuite, on construit l'opérateur $B = B^* > 0$ et l'on calcule les constantes γ_1, γ_2, enfin on cherche $n = n_0(\varepsilon)$ et les paramètres $\{\tau_k^*\}$.

S'il s'agit de la méthode triangulaire alternée avec opérateur factorisé $B = (\mathscr{D} + \omega R_1)\mathscr{D}^{-1}(\mathscr{D} + \omega R_2)$, il faut choisir la matrice $\mathscr{D}$ et les constantes δ, Δ (voir ch. X), connaissant δ et Δ, on détermine ω, γ_1, γ_2, etc.

On a fait appel dans le livre à un grand nombre d'exemples illustrant l'application des méthodes directes et itératives à la résolution des équations aux différences concrètes. En l'occurrence, dans le ch. XV on expose les méthodes de résolution des équations aux différences elliptiques en coordonnées curvilignes: cylindriques (r, z) et polaires (r, φ).

Dans le ch. XIV on étudie les problèmes multidimensionnels, les schémas pour équations de la théorie de l'élasticité, etc.

Il est important de noter que quelle que soit la méthode utilisée pour la résolution du problème de différences aux limites considéré, son traitement préalable s'effectue suivant le même organigramme: d'abord on construit l'opérateur A, ensuite il est étudié comme un opérateur dans l'espace H des fonctions de mailles. Une fois l'information sur le problème recueillie, on décide quelle méthode de résolution il faut choisir pour le problème en tenant compte de tous les aléas, y compris le type de machine utilisé, l'existence de programmes standards, etc.

CHAPITRE I

MÉTHODES DIRECTES DE RÉSOLUTION DES ÉQUATIONS AUX DIFFÉRENCES

On étudie dans ce chapitre la théorie générale des équations aux différences linéaires ainsi que les méthodes directes de résolution des équations avec coefficients constants fournissant la solution sous forme fermée. Dans le § 1 on présente des notions générales sur les équations de mailles. Le § 2 traite de la théorie générale des équations aux différences linéaires du m-ième ordre. Dans le § 3 sont exposées les méthodes de résolution des équations aux coefficients constants, tandis que dans le § 4 ces méthodes sont appliquées à la résolution des équations du deuxième ordre. Le § 5 fournit des solutions de problèmes discrets en valeurs propres pour le cas de l'opérateur de différences le plus simple.

§ 1. Equations de mailles. Notions générales

1. Maillages et fonctions de mailles. Un nombre important de problèmes de physique et d'applications techniques aboutissent à des équations différentielles aux dérivées partielles (aux équations de la physique mathématique). Les processus permanents de nature physique diverse se décrivent par des équations du type elliptique.

Les solutions précises des problèmes aux limites pour équations elliptiques ne s'obtiennent que dans des cas particuliers. Aussi ces problèmes sont-ils, en général, résolus de façon approchée. Une des méthodes universelles et efficaces ayant reçu une grande extension actuellement lorsqu'il s'agit de résoudre approximativement des équations de la physique mathématique est la méthode des différences finies.

L'essence de la méthode est la suivante. Le domaine de variation permanente de l'argument (par exemple, un segment, un rectangle, etc.) est remplacé par un ensemble discret de points (de nœuds) qu'on appelle *maillage* ou *réseau*. Au lieu de fonctions d'un argument continu on étudie les fonctions d'un argument discret définies aux nœuds du maillage et appelées *fonctions de mailles*. Les dérivées figurant dans les équations différentielles et les conditions aux limites sont remplacées par des différences divisées (rapports incrémentiels); en outre, le problème aux limites de l'équation différentielle est remplacé par un système d'équations algébriques linéai-

res ou non linéaires (équations de mailles ou équations aux différences). Ces systèmes sont souvent appelés *schémas aux différences.*

Arrêtons-nous plus en détail sur les notions de base de la méthode des différences finies. Voyons d'abord les exemples les plus simples de maillages.

E x e m p l e 1. *Maillage dans un domaine unidimensionnel.* Soit le segment $0 \leqslant x \leqslant l$ le domaine de variation de l'argument x. Divisons ce segment en N parties égales de longueur $h = l/N$ par les points $x_i = ih$, $i = 0, 1, \ldots, N$. L'ensemble de ces points est appelé *maillage régulier* sur le tronçon $[0, l]$ et est noté $\bar{\omega} = \{x_i = ih, i = 0, 1, \ldots, N, hN = l\}$, tandis que le nombre h est la distance entre les points (les nœuds) du maillage $\bar{\omega}$ appelée *pas du maillage.*

Pour la distinction d'une partie du maillage $\bar{\omega}$, on utilisera par la suite les notations suivantes

$$\omega = \{x_i = ih, \quad i = 1, 2, \ldots, N-1, \quad Nh = l\},$$
$$\omega^+ = \{x_i = ih, \quad i = 1, 2, \ldots, N, \quad Nh = l\},$$
$$\omega^- = \{x_i = ih, \quad i = 0, 1, \ldots, N-1, \quad Nh = l\},$$
$$\gamma = \{x_0 = 0, \quad x_N = l\}.$$

Le segment $[0, l]$ peut être divisé en N parties par introduction de points arbitraires $0 = x_0 < x_1 < \ldots < x_i < x_{i+1} < \ldots \ldots < x_{N-1} < x_N = l$. Dans ce cas on obtient un maillage $\bar{\omega} = \{x_i, i = 0, 1, \ldots, N, x_0 = 0, x_N = l\}$ de pas $h_i = x_i - x_{i-1}$ au nœud x_i, $i = 1, 2, \ldots, N$, qui est une fonction du numéro i du nœud x_i, autrement dit la fonction de maille $h_i = h(i)$.

Si $h_i \neq h_{i+1}$ pour au moins un numéro i, le maillage $\bar{\omega}$ est alors dit *irrégulier.* Si $h_i = h = l/N$, on obtient alors le maillage régulier construit plus haut. Pour un maillage irrégulier on introduit un pas moyen $\hbar_i = \hbar(i)$ au nœud x_i, $\hbar_i = 0{,}5(h_i + h_{i+1})$, $1 \leqslant i \leqslant N-1$, $\hbar_0 = 0{,}5h_1$, $\hbar_N = 0{,}5h_N$. Sur une droite infinie $-\infty < x < \infty$ on peut définir des maillages $\Omega = \{x_i = a + ih, i = 0, \pm 1, \pm 2, \ldots\}$ avec origine en un point quelconque $x = a$ et de pas h, possédant un nombre infini de nœuds.

E x e m p l e 2. *Maillage dans un domaine bidimensionnel.* Soit le rectangle $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha, \alpha = 1, 2\}$ de limite Γ le domaine de variation de l'argument $x = (x_1, x_2)$. Sur les segments $0 \leqslant x_\alpha \leqslant l_\alpha$ construisons des maillages réguliers $\bar{\omega}_\alpha$ de pas h_α:

$$\bar{\omega}_1 = \{x_1(i) = ih_1, \quad i = 0, 1, \ldots, M, \quad h_1 M = l_1\},$$
$$\bar{\omega}_2 = \{x_2(j) = jh_2, \quad j = 0, 1, \ldots, N, \quad h_2 N = l_2\}.$$

L'ensemble des nœuds $x_{ij} = (x_1(i), x_2(j))$ possédant des coordonnées dans le plan $x_1(i)$ et $x_2(j)$ est appelé *maillage sur le rectangle* $\overline{G}$ et est noté par $\overline{\omega} = \{x_{ij} = (ih_1, jh_2), i = 0, 1, \ldots, M, j = 0, 1, \ldots, N, h_1M = l_1, h_2N = l_2\}$.

Le maillage $\overline{\omega}$ est apparemment composé de points d'intersection des droites $x_1 = x_1(i)$ et $x_2 = x_2(j)$.

Le maillage ainsi construit est régulier par rapport à chacune des variables x_1 et x_2. Si l'un au moins des maillages $\overline{\omega}_\alpha$ est irrégulier, le maillage $\overline{\omega}$ est alors dit *irrégulier*. Si $h_1 = h_2$, le maillage est dit *carré* (à mailles carrées), dans le cas contraire il est *rectangle*.

Les points de $\overline{\omega}$ appartenant à Γ sont appelés *points frontières* et leur ensemble constitue la frontière du maillage: $\gamma = \{x_{ij} \in \Gamma\}$.

Il est commode pour décrire la structure du maillage $\overline{\omega}$ d'utiliser l'écriture $\overline{\omega} = \overline{\omega}_1 \times \overline{\omega}_2$ en représentant $\overline{\omega}$ comme un produit topologique des maillages $\overline{\omega}_1$ et $\overline{\omega}_2$. En se servant des notations ω^+, ω^- et ω introduites dans l'exemple 1, on peut isoler des parties du maillage rectangulaire $\overline{\omega}$, par exemple:

$$\omega_1 \times \omega_2^+ = \{x_{ij} = (ih_1, jh_2), \quad i = 1, 2, \ldots, M-1, \quad j = 1, 2, \ldots, N\},$$

$$\omega_1^- \times \overline{\omega}_2 = \{x_{ij} = (ih_1, jh_2), \quad i = 0, 1, \ldots, M-1, \quad j = 0, 1, \ldots, N\}.$$

Etudions maintenant la notion de fonction de maille. Soit $\overline{\omega}$ le maillage introduit dans le domaine unidimensionnel, x_i étant les nœuds du maillage. La fonction $y = y(x_i)$ de l'argument discret x_i est appelée *fonction de maille* définie sur le maillage $\overline{\omega}$. De façon analogue se détermine la fonction de maille de tout maillage $\overline{\omega}$ introduit dans le domaine de variation de l'argument continu. Par exemple, si x_{ij} est un nœud du maillage $\overline{\omega}$ d'un domaine bidimensionnel, alors $y = y(x_{ij})$. Il est évident que les fonctions de mailles peuvent aussi être assimilées à des fonctions d'argument entier représentant le numéro du nœud dans le maillage. C'est ainsi qu'on peut écrire $y = y(x_i) = y(i)$, $y = y(x_{ij}) = y(i, j)$. Quelquefois, pour désigner les fonctions de mailles on utilisera l'écriture suivante: $y(x_i) = y_i$, $y(x_{ij}) = y_{ij}$.

La fonction de maille y_i peut se représenter sous forme de vecteur en considérant les valeurs de la fonction comme des composantes du vecteur $Y = (y_0, y_1, \ldots, y_N)$. Dans cet exemple y_i est défini sur le maillage $\overline{\omega} = \{x_i, i = 0, 1, \ldots, N\}$, comprenant $N + 1$ nœuds, tandis que le vecteur Y a la dimension $N + 1$. Si $\overline{\omega}$ est un maillage dans un rectangle $(\overline{\omega} = \{x_{ij} = (ih_1, jh_2), i = 0, 1, \ldots$

$\ldots, M$, $j = 0, 1, \ldots, N\}$), à la fonction de maille y_{ij} traduite sur $\bar{\omega}$ correspond alors le vecteur $Y = (y_{00}, \ldots, y_{M0}, y_{01}, \ldots, y_{M1}, \ldots, y_{0N}, \ldots, y_{MN})$ de dimension $(M + 1)(N + 1)$. Les nœuds du maillage $\bar{\omega}$ sont dans ce cas considérés comme ordonnés suivant les lignes du maillage.

On a examiné les fonctions de mailles scalaires, c'est-à-dire les fonctions dont les valeurs en chaque nœud du maillage sont des nombres. Donnons maintenant des exemples de *fonctions de mailles vectorielles*, dont les valeurs au nœud sont des vecteurs. Si dans l'exemple pris plus haut on désigne par $Y(x_2(j)) = Y_j$ le vecteur dont les composantes sont des valeurs de la fonction de maille y_{ij} aux nœuds $x_{0j}, x_{1j}, \ldots, x_{Mj}$ de la j-ième ligne du maillage $\bar{\omega}$: $Y_j = (y_{0j}, y_{1j}, \ldots, y_{Mj})$, $j = 0, 1, \ldots, N$, on obtiendra alors la fonction de maille vectorielle Y_j donnée sur le maillage $\bar{\omega}_2 = \{x_2(j) = jh_2,\ j = 0, 1, \ldots, N\}$.

Si la fonction donnée sur le maillage prend des valeurs complexes, cette fonction de maille est dénommée *complexe*.

2. Différences divisées et quelques identités de différences. Soit un maillage $\bar{\omega}$. L'ensemble de toutes les fonctions de mailles associées au maillage $\bar{\omega}$ est un espace vectoriel obéissant à des règles déterminées d'addition et de multiplication des fonctions par un nombre. Sur l'espace des fonctions de mailles on peut définir des opérateurs de différences ou des opérateurs discrets. L'opérateur Λ transformant la fonction de maille y en fonction de maille $f = \Lambda y$ est appelé *opérateur de différences* ou *discret*. L'ensemble des nœuds du maillage utilisé lors de l'écriture de l'opérateur de différences au nœud du maillage est appelé *stencil* (ou support) de cet opérateur.

L'opérateur de différences le plus simple est l'opérateur de différentiation au sens des différences finies d'une fonction de maille engendrant des différences divisées (des approximations au sens des différences finies). Définissons ces différences.

Soit Ω un maillage régulier de pas h donné sur la droite $-\infty < x < \infty$: $\Omega = \{x_i = a + ih,\ i = 0, \pm 1, \pm 2, \ldots\}$. Les différences premières se définissent pour la fonction de maille $y_i = y(x_i)$, $x_i \in \Omega$ à l'aide des formules

$$\Lambda_1 y_i = y_{\bar{x},i} = \frac{y_i - y_{i-1}}{h}, \quad \Lambda_2 y_i = y_{x,i} = \frac{y_{i+1} - y_i}{h} \tag{1}$$

et s'appellent respectivement *différences progressive* et *régressive*. On utilise également la *différence centrale*

$$\Lambda_3 y_i = y_{\overset{\circ}{x},i} = \frac{y_{i+1} - y_{i-1}}{2h} = 0{,}5(\Lambda_1 + \Lambda_2)\, y_i. \tag{2}$$

Si le maillage est irrégulier, on utilise pour les différences premières les notations suivantes:

$$y_{\bar{x},i}=\frac{y_i-y_{i-1}}{h_i},\quad y_{x,i}=\frac{y_{i+1}-y_i}{h_{i+1}},\quad y_{\hat{x},i}=\frac{y_{i+1}-y_i}{\hbar_i},\tag{3}$$

$$y_{\overset{\circ}{x},i}=0{,}5\,(y_{\bar{x},i}+y_{x,i}),\quad \hbar_i=0{,}5\,(h_i+h_{i+1}).$$

Il s'ensuit des définitions (1) et (3) les relations suivantes:

$$y_{x,i}=y_{\bar{x},i+1},\tag{4}$$

$$y_{x,i}=\frac{\hbar_i}{h_{i+1}}y_{\hat{x},i},\tag{5}$$

ainsi que les égalités

$$y_i=y_{i+1}-h_{i+1}y_{x,i}=y_{i-1}+h_iy_{\bar{x},i}.\tag{6}$$

Les opérateurs de différences Λ_1, Λ_2 et Λ_3 présentent des stencils composés de deux points et sont utilisés pour l'approximation de la dérivée première $Lu=u'$ de la fonction $u=u(x)$ à une variable. Les opérateurs Λ_1 et Λ_2 approximant l'opérateur L sur les fonctions lisses avec l'erreur $O(h)$, et Λ_3 — avec l'erreur $O(h^2)$.

Les *différences d'ordre n* se définissent comme des fonctions de mailles obtenues par calcul de la différence première de la fonction constituant une différence d'ordre $n-1$. Donnons quelques exemples de différences secondes:

$$y_{\bar{x}x,i}=\frac{y_{\bar{x},i+1}-y_{\bar{x},i}}{h}=\frac{1}{h^2}(y_{i-1}-2y_i+y_{i+1}),$$

$$y_{\overset{\circ}{x}\overset{\circ}{x},i}=\frac{y_{\overset{\circ}{x},i+1}-y_{\overset{\circ}{x},i-1}}{2h}=\frac{1}{4h^2}(y_{i-2}-2y_i+y_{i+2}),$$

$$y_{\bar{x}\hat{x},i}=\frac{1}{\hbar_i}(y_{\bar{x},i+1}-y_{\bar{x},i})=\frac{1}{\hbar_i}(y_{x,i}-y_{\bar{x},i})=$$

$$=\frac{1}{\hbar_i}\left(\frac{y_{i+1}-y_i}{h_{i+1}}-\frac{y_i-y_{i-1}}{h_i}\right),$$

utilisées lors de l'approximation de la dérivée seconde $Lu=u''$ de la fonction $u=u(x)$. Dans le cas d'un maillage régulier l'erreur d'approximation vaut $O(h^2)$. Les opérateurs de différences correspondants possèdent un stencil triponctuel. Dans l'approximation de la dérivée d'ordre 4 $Lu=u^{IV}$ on se sert de la différence d'ordre 4 $y_{\bar{x}x\bar{x}x,i}=\frac{1}{h^4}(y_{i-2}-4y_{i-1}+6y_i-4y_{i+1}+y_{i+2})$. De façon analogue, dans l'approximation des dérivées d'ordre n on recourt aux différences d'ordre n.

Il n'est pas très difficile de déterminer les différences dans le cas des fonctions de mailles à plusieurs variables.

Pour transformer les expressions renfermant les différences des fonctions de mailles on doit recourir aux formules de dérivation au sens de différences finies d'un produit de fonctions de mailles et à celles de sommation par parties. Ces formules sont analogues aux formules du calcul différentiel.

1) *Formules de dérivation de produit au sens de différences finies.* En utilisant les définitions des différences (3), il est aisé de vérifier qu'on a les identités:

$$(uv)_{\bar{x},\,i} = u_{\bar{x},\,i}v_{i-1} + u_i v_{\bar{x},\,i} = u_{\bar{x},\,i}v_i + u_{i-1}v_{\bar{x},\,i} = u_{\bar{x},\,i}v_i + u_i v_{\bar{x},\,i} - h_i u_{\bar{x},\,i}v_{\bar{x},\,i},$$

$$(uv)_{x,\,i} = u_{x,\,i}v_{i+1} + u_i v_{x,\,i} = u_{x,\,i}v_i + u_{i+1}v_{x,\,i} = u_{x,\,i}v_i + u_i v_{x,\,i} + h_{i+1}u_{x,\,i}v_{x,\,i},$$

$$(uv)_{\hat{x},\,i} = u_{\hat{x},\,i}v_{i+1} + u_i v_{\hat{x},\,i} = u_{\hat{x},\,i}v_i + u_{i+1}v_{\hat{x},\,i} = u_{\hat{x},\,i}v_i + u_i v_{\hat{x},\,i} + \hbar_i u_{\hat{x},\,i}v_{\hat{x},\,i}.$$

En recourant à (4), (5), cette dernière identité peut être récrite sous la forme

$$(uv)_{\hat{x},\,i} = u_{\hat{x},\,i}v_i + \frac{h_{i+1}}{\hbar_i}u_{i+1}v_{\bar{x},\,i+1}. \tag{7}$$

2) *Formules de sommation par parties.* En multipliant (7) par $\hbar_i$ et en sommant le rapport obtenu en i de $m+1$ à $n-1$, on obtient:

$$\sum_{i=m+1}^{n-1}(uv)_{\hat{x},\,i}\hbar_i = u_n v_n - u_{m+1}v_{m+1} = \sum_{i=m+1}^{n-1}u_{\hat{x},\,i}v_i\hbar_i + \sum_{i=m+1}^{n-1}u_{i+1}v_{\bar{x},\,i+1}h_{i+1}.$$

Profitant de (6), on obtient la relation $v_{m+1} = v_m + h_{m+1}v_{x,\,m} = v_m + h_{m+1}v_{\bar{x},m+1}$, qu'on porte dans l'égalité trouvée plus haut. Il vient finalement

$$u_n v_n - u_{m+1}v_m = \sum_{i=m+1}^{n-1}u_{\hat{x},\,i}v_i\hbar_i + \sum_{i=m}^{n-1}u_{i+1}v_{\bar{x},\,i+1}h_{i+1}.$$

La substitution de l'indice de sommation $i' = i - 1$ dans la seconde somme du deuxième membre fournit la formule suivante de sommation par parties:

$$\sum_{i=m+1}^{n-1}u_{\hat{x},\,i}v_i\hbar_i = -\sum_{i=m+1}^{n}u_i v_{\bar{x},\,i}h_i + u_n v_n - u_{m+1}v_m. \tag{8}$$

En utilisant (6), on obtient sans peine à partir de (8) encore une formule de sommation par parties

$$\sum_{i=m+1}^{n-1}u_{\bar{x},\,i}v_i h_i = -\sum_{i=m}^{n-1}u_i v_{\hat{x},\,i}\hbar_i + u_{n-1}v_n - u_m v_m. \tag{9}$$

De la formule (8) il s'ensuit que la fonction u_i doit être définie pour $m+1 \leqslant i \leqslant n$, et la fonction v_i pour $m \leqslant i \leqslant n$. Soit maintenant y_i une fonction de maille définie pour $m \leqslant i \leqslant n$. La fonction $u_i = y_{\bar{x}, i}$ est alors définie pour $m+1 \leqslant i \leqslant n$. En portant u_i dans (8) on obtient l'identité suivante:

$$\sum_{i=m+1}^{n-1} y_{\bar{x}\hat{x}, i} v_i \hbar_i = -\sum_{i=m+1}^{n} y_{\bar{x}, i} v_{\bar{x}, i} h_i + y_{\bar{x}, n} v_n - y_{x, m} v_m. \quad (10)$$

On a le

Lemme 1. *Soit donnée sur un maillage irrégulier quelconque* $\bar{\omega} = \{x_i,\ i = 0, 1, \ldots, N,\ x_0 = 0,\ x_N = l\}$ *une fonction de maille* y_i *s'annulant pour* $i = 0$, $i = N$. *Cette fonction implique l'égalité*

$$\sum_{i=1}^{N-1} y_{\bar{x}\hat{x}, i} y_i \hbar_i = -\sum_{i=1}^{N} (y_{\bar{x}, i})^2 h_i.$$

Le lemme 1 s'ensuit de façon évidente de l'identité (10).

Corollaire. *Si* $\bar{\omega}$ *est un maillage régulier,* $y_0 = y_N = 0$ *et* $y_i \not\equiv 0$, *alors* $\sum_{i=1}^{N-1} y_{\bar{x}x, i} y_i h = -\sum_{i=1}^{N} y^2_{\bar{x}, i} h < 0$.

L'étude des formules de différences finies s'arrête à ce point. D'autres formules seront examinées au chapitre V.

Les identités obtenues sont utilisées non seulement pour la transformation des expressions aux différences finies. Elles sont souvent appliquées, par exemple, pour le calcul de différentes sommes et séries finies.

Donnons un exemple. Il s'agit de calculer la somme $S_n = \sum_{i=1}^{n-1} i a^i$, $a \neq 1$. Introduisons les fonctions de mailles suivantes données sur un maillage régulier $\bar{\omega} = \{x_i = i,\ i = 0, 1, \ldots, N,\ h = 1\}$:

$$v_i = i, \quad u_i = (a^i - a^n)/(a-1). \quad (11)$$

Sur le maillage impliqué la formule de sommation par parties (8) pour toutes fonctions de mailles prend la forme ($m = 0$)

$$\sum_{i=1}^{n-1} u_{x, i} v_i = -\sum_{i=1}^{n} u_i v_{\bar{x}, i} + u_n v_n - u_1 v_0.$$

Compte tenu de ce que pour les fonctions (11) se vérifient les relations $v_0 = u_n = 0$, $v_{\bar{x}, i} = 1$, $u_{x, i} = a^i$, il vient

$$S_n = \sum_{i=1}^{n-1} i a^i = -\sum_{i=1}^{n} \frac{a^i - a^n}{a-1} = \frac{a^n(n(a-1) - a) + a}{(a-1)^2}.$$

La somme cherchée est trouvée.

3. Les équations de mailles et aux différences finies. Soit $y_i = y(i)$ la fonction de maille d'un argument discret i. De son côté, la valeur de la fonction de maille $y(i)$ constitue un ensemble discret. On peut sur cet ensemble définir la fonction de maille qui, une fois égalée à zéro, fournit l'équation de la fonction de maille $y(i)$, appelée *équation de maille*. L'*équation aux différences* est un cas particulier de l'équation de maille. L'objet de nos études se portera essentiellement sur les équations aux différences.

On obtient des équations de mailles en approximant sur un maillage les équations intégrales et différentielles.

Donnons tout d'abord des exemples d'approximations au sens de différences finies des équations différentielles ordinaires.

C'est ainsi que les équations différentielles du premier ordre $\frac{du}{dx} = f(x)$, $x > 0$, sont remplacées par des équations aux différences d'ordre un $\frac{y_{i+1} - y_i}{h} = f(x_i)$, $x_i = ih$, $i = 0, 1, \ldots$ ou $y_{i+1} = y_i + hf(x_i)$, où h est le pas du maillage $\omega = \{x_i = ih, i = 0, 1, \ldots\}$. La fonction cherchée est la fonction de maille $y_i = y(i)$.

Dans l'approximation au sens de différences finies de l'équation du deuxième ordre $\frac{d^2u}{dx^2} = f(x)$ on obtient une équation aux différences d'ordre deux $y_{i+1} - 2y_i + y_{i-1} = \varphi_i$, $\varphi_i = h^2 f_i$, $f_i = f(x_i)$, $x_i = ih$. Si l'approximation est réalisée sur un stencil triponctuel (x_{i-1}, x_i, x_{i+1}) d'une équation de forme générale $(ku')' + ru' - qu = f(x)$, on obtient une équation aux différences d'ordre deux aux coefficients variables de la forme $a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -\varphi_i$, $i = 0, 1, \ldots$, où a_i, c_i, b_i, φ_i sont des fonctions de mailles données, tandis que y_i est la fonction de maille cherchée.

L'approximation sur maillage d'une équation du quatrième ordre $(ku'')'' = f(x)$ aboutit à une équation aux différences d'ordre quatre; sa forme est:

$$a_i^{(2)} y_{i-2} + a_i^{(1)} y_{i-1} + c_i y_i + b_i^{(1)} y_{i+1} + b_i^{(2)} y_{i+2} = \varphi_i.$$

Pour l'approximation au sens de différences finies des dérivées u', u'', u''' on peut utiliser des stencils possédant un grand nombre de nœuds. On aboutit ainsi à des équations aux différences d'un ordre plus élevé.

L'équation linéaire associée à la fonction de maille $y(i)$ (fonction de l'argument entier i)

$$a_0(i)\, y(i) + a_1(i)\, y(i+1) + \ldots + a_m(i)\, y(i+m) = f(i), \quad (12)$$

où $a_0(i) \neq 0$, $a_m(i) \neq 0$ et $f(i)$ une fonction de maille donnée, est appelée *équation aux différences d'ordre m*.

Si (12) ne contient pas $y(i)$ mais contient $y(i+1)$, la substitution de la variable indépendante i' à $i+1$ transforme cette équation en l'équation d'ordre $m-1$.

C'est en quoi réside une des différences des équations de mailles vis-à-vis des équations différentielles, où la substitution de la variable indépendante n'engendre pas de changement d'ordre dans l'équation.

Soit $F(i, y(i), y(i+1), \ldots, y(i+m))$ une fonction de maille non linéaire. Alors $F(i, y(i), y(i+1), \ldots, y(i+m)) = 0$ est *une équation aux différences non linéaire d'ordre* m, si F dépend explicitement de $y(i)$ et $y(i+m)$.

Pour faciliter la comparaison avec les équations différentielles, introduisons *les différences* (*progressives*) *pour les fonctions de mailles*: $\Delta y_i = y_{i+1} - y_i$, $\Delta^2 y_i = \Delta(\Delta y_i), \ldots, \Delta^{k+1} y_i = \Delta(\Delta^k y_i)$, $k = 1, 2, \ldots$

On peut alors récrire (12) sous la forme

$$\alpha_0(i)\, y(i) + \alpha_1(i)\, \Delta y_i + \ldots + \alpha_m(i)\, \Delta^m y_i = f_i, \qquad (12')$$

où $\alpha_m(i) = a_m(i) \neq 0$. De plus, le coefficient a_0 associé à y_0 est également différent de zéro.

L'équation aux différences (12') est l'analogue formel de l'équation différentielle d'ordre m:

$$\alpha_0 u + \alpha_1 \frac{du}{dx} + \ldots + \alpha_{m-1} \frac{d^{m-1}u}{dx^{m-1}} + \alpha_m \frac{d^m u}{dx^m} = f(x),$$

où $\alpha_m \neq 0$, $\alpha_k = \alpha_k(x)$, $k = 0, 1, \ldots, m$. Soit un maillage $\omega = \{x_i = ih,\ i = 0, 1, \ldots\}$. En posant

$$y_{x,i} = \frac{y_{i+1} - y_i}{h}, \qquad y_{xx,i} = (y_x)_{x,i}, \ldots, \; y_x^{(k)} = y_{\underbrace{x \ldots x}_{k \text{ fois}},\, i},$$

de manière que $y_x^{(k)} = (y_x^{(k-1)})_x$, $k \geqslant 1$, $y_{x,i}^{(0)} = y(i)$, on exprimera $y(i+k)$ en fonction de $y(i), y_x^{(1)}, \ldots, y_x^{(k-1)}$, ainsi, par exemple, $y(i+3) = y(i) + 3hy_{x,i} + 3h^2 y_{xx,i} + h^3 y_{xxx,i}$.

L'équation (12) s'écrira alors sous la forme

$$\bar{\alpha}_0 y(i) + \bar{\alpha}_1(i)\, y_x(i) + \ldots + \bar{\alpha}_{m-1} y_x^{(m-1)}(i) + \bar{\alpha}_m y_x^{(m)}(i) = f_i,$$

où $\bar{\alpha}_m = a_m \neq 0$ et $a_0 \neq 0$. L'analogie avec l'équation différentielle d'ordre m est ici évidente.

De façon analogue se détermine l'équation aux différences associée à la fonction de maille $y_{i_1, i_2} = y(i_1, i_2)$ à deux arguments discrets et, en général, à un nombre quelconque d'arguments. Par exemple, le schéma pentaponctuel « croix » de l'équation de Poisson $\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -f(x_1, x_2)$ sur maillage $\omega = \{x_i =$

$= (i_1h_1,\ i_2h_2),\ i_1,\ i_2 = 0,\ 1,\ \ldots\}$ a la forme

$$\frac{y(i_1-1,\ i_2)-2y(i_1,\ i_2)+y(i_1+1,\ i_2)}{h_1^2}+$$
$$+\frac{y(i_1,\ i_2-1)-2y(i_1,\ i_2)+y(i_1,\ i_2+1)}{h_2^2}=f_{i_1i_2}$$

et constitue une équation aux différences d'ordre deux par rapport à chacun des arguments discrets i_1 et i_2.

On obtient une équation de maille de forme *générale* en approximant l'équation intégrale $u(x) = \int\limits_0^1 K(x,\ s)\,u(s)\,ds + f(x),\ 0 \leqslant x \leqslant 1$ sur maillage $\bar\omega = \{x_i = ih,\ i = 0.\ 1.\ \ldots.\ N,\ hN = 1\}$. Substituons à l'intégrale la somme

$$\int\limits_0^1 K(x,\ s)\,u(s)\,ds \approx h\sum_{j=0}^{N}\alpha_jK(x,\ jh)\,u(jh),$$

où α_j est un coefficient de quadrature, et à la place de l'équation intégrale écrivons l'équation de maille

$$y_i=\sum_{j=0}^{N}\alpha_jK(ih,\ jh)\,y_j+f_i,\quad i=0,\ 1,\ \ldots,\ N,$$

où la sommation s'effectue par rapport à tous les nœuds du maillage $\bar\omega$, l'inconnue étant la fonction y_i.

L'équation de maille peut s'écrire sous la forme

$$\sum_{j=0}^{N}c_{ij}y_j=f_i,\quad i=0,\ 1,\ \ldots,\ N. \tag{13}$$

Elle renferme toutes les valeurs $y_0,\ y_1,\ \ldots.\ y_N$ de la fonction de maille. On peut l'assimiler à une équation aux différences d'ordre N égal au nombre de nœuds du maillage moins un.

L'équation aux différences (12) d'ordre m est un cas spécial de l'équation de maille correspondant au cas où la matrice (c_{ij}) ne possède des éléments non nuls que sur les diagonales m parallèles à la principale diagonale.

Dans le cas général il est sous-entendu que i est non seulement l'indice $i = 0,\ 1, \ldots$ mais aussi le multiindice. c.-à-d. le vecteur $i = (i_1,\ i_2,\ \ldots.\ i_p)$ aux composantes entières $i_\alpha = 0,\ 1,\ 2,\ \ldots,$ $\alpha = 1,\ 2,\ \ldots.\ p$, avec $i \in \omega$, où ω est le maillage.

L'équation de maille du type linéaire a la forme

$$\sum_{j\in\omega}c_{ij}y_j=f_i,\quad i\in\omega, \tag{14}$$

où la sommation s'effectue en tous les nœuds du maillage ω, f_i étant la fonction de maille donnée et y_i la fonction de maille cherchée.

Si l'on numérote tous les nœuds du maillage, on peut alors écrire $y_i = y(i)$, où i est le numéro de nœud, $i = 0, 1, 2, \ldots, N$. L'équation de maille (14) acquerra dans ce cas la forme (13).

Il va de soi que c'est un système d'équations algébriques linéaires d'ordre $N + 1$ à matrice (c_{ij}). Donc, tout système d'équations algébriques linéaires peut être traité comme une équation de maille et inversement.

Si $y(i)$ est une fonction de maille *vectorielle*, on dit alors qu'on a affaire à une *équation de maille (discrète) vectorielle d'ordre m.*

Soit $F(i, y_0, y_1, \ldots, y_N)$ une fonction donnée (en général, non linéaire) de $N + 2$ arguments $i, y_0, y_1, \ldots, y_N$. En l'égalant à zéro, on obtient l'équation de maille non linéaire $F(i, y_0, y_1, \ldots$ $\ldots, y_N) = 0$, $i = 0, 1, \ldots, N$, dont la solution est appelée fonction de maille $y(i)$ transformant cette équation en une identité.

Etudions la fonction de maille $\mathcal{F}(i) = F(i, y_0, y_1, \ldots, y_N)$, $i = 0, 1, \ldots, N$. On voit que la fonction F définit un certain opérateur de maille (discret) qui transforme la fonction de maille $y(i)$ en fonction de maille $\mathcal{F}(i)$.

Si F est une fonction linéaire, on obtient l'équation (14) qui, apparemment, peut être écrite sous forme opératorielle $Ay = f$, où A est un opérateur linéaire à matrice (c_{ij}) et y un vecteur dans l'espace des fonctions de mailles.

Si les coefficients c_{ij} ne dépendent pas de j, (14) est appelé *équation de maille à coefficients constants.*

Bien que dans ce livre l'attention soit essentiellement pointée sur la résolution numérique des équations aux différences obtenues par approximation au sens de différences finies d'équations différentielles du type elliptique, les méthodes itératives s'appliquent à toute équation de maille linéaire, c'est-à-dire à tout système d'équations algébriques linéaires. Aussi les éléments théoriques sur les méthodes itératives exposées dans cet ouvrage sont-ils de nature générale. Le caractère spécifique des équations de mailles est l'ordre élevé de ce système, l'ordre de l'équation augmentant avec le resserrement des mailles (le nombre d'inconnues est égal au nombre N de nœuds du maillage, $N = O\left(\frac{1}{h^p}\right)$ au cas de p composantes, h étant le pas du maillage).

4. Problème de Cauchy et problèmes aux limites pour équations aux différences. Donnons quelques exemples supplémentaires d'équations aux différences en nous arrêtant sur la position des problèmes pour les équations aux différences.

Notons que les exemples les plus simples d'équations aux différences d'ordre un sont les formules pour les termes de progressions

arithmétique et géométrique:

$$y_{i+1} = y_i + d, \quad y_{i+1} = qy_i, \quad i = 0, 1, \ldots$$

La solution de l'équation du premier ordre peut être obtenue si sont données les conditions initiales pour $i = 0$ (problème de Cauchy).

La solution $y(i + m)$ de l'équation aux différences d'ordre m est complètement déterminée par les valeurs $y(i)$ définies en m points quelconques disposés successivement $i_0, i_0 + 1, \ldots, i_0 + m - 1$. Et de fait, puisque $a_m(i) \neq 0$, de (12) on tire $y(i + m) = b_{m-1}(i)\, y(i + m - 1) + \ldots + b_0(i)\, y(i) + \varphi(i)$. En posant successivement $i = i_0, i_0 + 1, \ldots$, on obtient les valeurs $y(i)$ pour $i \geqslant i_0$. De façon analogue, en exprimant sur la base de (12) $y(i)$ en fonction de $y(i + 1), \ldots, y(i + m)$ et en posant successivement $i = i_0 - 1, i_0 - 2, \ldots$, on aboutit à $y(i)$ pour $i \leqslant i_0 - 1$. S'il s'agit dans l'équation (12) de déterminer $y(i)$ pour $i \geqslant 0$, il suffit de définir la valeur en m nœuds (conditions initiales) $y(0) = y_0$, $y(1) = y_1, \ldots, y(m - 1) = y_{m-1}$.

En joignant ces conditions à l'équation (12), on obtient le problème de Cauchy ou le problème de conditions initiales pour l'équation aux différences d'ordre m.

Pour les équations du premier ordre ($m = 1$), il suffit, comme on l'a vu, de fixer une condition initiale.

On obtient les équations aux différences non linéaires quand on résout des équations différentielles non linéaires. Voyons par exemple l'équation différentielle

$$\frac{du}{dx} = f(x, u), \quad x > 0, \quad u(0) = \mu_1$$

(problème de Cauchy). En lui substituant le schéma d'Euler (schéma explicite), on obtient l'équation aux différences d'ordre un $y_{i+1} = y_i + hf(x_i, y_i)$, $i \geqslant 0$, $y_0 = \mu_1$.

Si à la dérivée du/dx on substitue pour $x = x_i = ih$ la différence régressive, on obtient l'équation aux différences non linéaire d'ordre un en y_i: $y_i = y_{i-1} + hf(x_i, y_i)$, $i > 0$, $y_0 = \mu_1$. Pour déterminer y_i il faut résoudre l'équation non linéaire $\varphi(y_i) = y_i - hf(x_i, y_i) = y_{i-1}$.

Etudions maintenant un exemple d'équation aux différences d'ordre deux. Posons qu'il s'agit de calculer les intégrales

$$I_k(\varphi) = \int_0^{\pi} \frac{\cos k\psi - \cos k\varphi}{\cos \psi - \cos \varphi}\, d\psi, \quad k = 0, 1, 2, \ldots$$

Notons avant tout que $I_0(\varphi) = 0$, $I_1(\varphi) = \pi$. Transformons l'expression $[\cos(k + 1)\psi - \cos(k + 1)\varphi] + [\cos(k - 1)\psi - \cos(k - 1)\varphi] = 2\cos k\psi \cos\psi - 2\cos k\varphi \cos\varphi = 2(\cos k\psi - \cos k\varphi)\cos\varphi + 2(\cos\psi - \cos\varphi)\cos k\psi$. Une fois celle-ci uti-

lisée, il vient

$$I_{k+1}(\varphi)+I_{k-1}(\varphi)=2\cos\varphi I_k(\varphi)+2\int_0^{\pi}\cos k\psi\, d\psi=2\cos\varphi I_k(\varphi),$$
$$k\geqslant 1.$$

Le calcul des intégrales $I_k(\varphi)$ se réduit donc à la résolution du problème de Cauchy pour le cas d'une équation aux différences d'ordre deux

$$I_{k+1}(\varphi)-2\cos\varphi I_k(\varphi)+I_{k-1}(\varphi)=0,\quad k\geqslant 1,\quad I_0(\varphi)=0,\quad I_1(\varphi)=\pi. \tag{15}$$

Examinons encore un exemple. Il s'agit de trouver la solution d'un problème aux limites pour un système d'équations différentielles ordinaires d'ordre un

$$\frac{d\mathbf{u}}{dx}=A\mathbf{u}+\mathbf{f}(x),\quad 0<x<l, \tag{16}$$

$B\mathbf{u}=\boldsymbol{\mu}_1$ pour $x=0$, $C\mathbf{u}=\boldsymbol{\mu}_2$ pour $x=l$. $\mathbf{u}(x)=(u_1(x), u_2(x), \ldots, u_M(x))$ est ici une fonction-vecteur de dimension M, $A=A(x)$ une matrice carrée de dimension $M\times M$, B et C des matrices rectangulaires de dimension $M_1\times M$ et $M_2\times M$ respectivement, $M_1+M_2=M$. Les vecteurs $\mathbf{f}(x)$, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ sont donnés et ont pour dimension M, M_1 et M_2 respectivement.

En introduisant sur le segment $0\leqslant x\leqslant l$ un maillage régulier $\bar\omega=\{x_i=ih,\ i=0, 1, \ldots, N,\ h=l/N\}$ et en définissant sur ce dernier la fonction de maille vectorielle $\mathbf{Y}_i=(y_1(i), y_2(i), \ldots \ldots, y_M(i))$, accordons avec le problème (16) le schéma aux différences du type le plus simple

$$\begin{gathered}\mathbf{Y}_{i+1}-(E+hA_i)\mathbf{Y}_i=\mathbf{F}_i,\quad 0\leqslant i\leqslant N-1,\\ B\mathbf{Y}_0=\boldsymbol{\mu}_1,\quad C\mathbf{Y}_N=\boldsymbol{\mu}_2,\end{gathered} \tag{17}$$

où $\mathbf{F}_i=h\mathbf{f}(x_i)$. C'est un exemple d'équation aux différences linéaire vectorielle d'ordre un à M_1 conditions pour $i=0$ et M_2 conditions pour $i=N$. On a ainsi pour un système d'équations aux différences d'ordre un le problème aux limites.

Pour les équations du second ordre les problèmes aux limites sont les plus typiques. Voyons, par exemple, le premier problème aux limites

$$\begin{gathered}\frac{d^2u}{dx^2}-q(x)u=-f(x),\quad 0<x<l,\quad u(0)=\mu_1,\\ u(l)=\mu_2,\quad q(x)\geqslant 0.\end{gathered} \tag{18}$$

Choisissons comme maillage $\bar\omega=\{x_i=ih,\ i=0, 1, \ldots, N,\ h= =l/N\}$ et posons au problème (18) en conséquence le problème aux

limites au sens de différences finies

$$y_{\bar{x}x,\,i} - d_i y_i = -\varphi_i, \quad 0 < i < N, \quad y_0 = \mu_1, \quad y_N = \mu_2, \tag{19}$$

où $d_i = q(x_i)$, $\varphi_i = f(x_i)$ pour des $q(x)$, $f(x)$ lisses. Ce problème constitue un cas particulier du problème aux limites pour une équation aux différences d'ordre deux

$$\begin{gathered} -a_i y_{i-1} + c_i y_i - b_i y_{i+1} = \varphi_i, \quad 1 \leqslant i \leqslant N-1, \\ y_0 = \mu_1, \quad y_N = \mu_2 \end{gathered} \tag{20}$$

avec $a_i = b_i = 1/h^2$, $c_i = d_i + 2/h^2$.

Le problème de différences (20) peut être écrit sous la forme

$$\mathcal{A} Y = F, \tag{21}$$

où $Y = (y_1, y_2, \ldots, y_{N-1})$ est un vecteur inconnu, $F = \left(\varphi_1 + \frac{1}{h^2}\mu_1, \varphi_2, \ldots, \varphi_{N-2}, \varphi_{N-1} + \frac{1}{h^2}\mu_2\right)$ un vecteur connu de dimension $N-1$, $\mathcal{A}$ une matrice carrée tridiagonale de la forme

$$\mathcal{A} = \left\| \begin{matrix} c_1 & -b_1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ -a_2 & c_2 & -b_2 & 0 & \ldots & 0 & 0 & 0 \\ 0 & -a_3 & c_3 & -b_3 & \ldots & 0 & 0 & 0 \\ . & . & . & . & \ldots & . & . & . \\ 0 & 0 & 0 & 0 & \ldots & c_{N-3} & -b_{N-3} & 0 \\ 0 & 0 & 0 & 0 & \ldots & -a_{N-2} & c_{N-2} & -b_{N-2} \\ 0 & 0 & 0 & 0 & \ldots & 0 & -a_{N-1} & c_{N-1} \end{matrix} \right\| . \tag{22}$$

On voit aussitôt que le problème aux limites pour l'équation aux différences d'ordre deux (20) constitue un système d'équations algébriques linéaires d'aspect spécial. Si le problème de Cauchy pour une équation aux différences d'ordre deux a toujours une solution, le premier problème aux limites (20) n'admet une solution pour toute partie droite qu'au cas où la matrice $\mathcal{A}$ du système (21) n'est pas dégénérée.

Les problèmes aux limites pour des équations aux différences d'ordre m aboutissent à des systèmes d'équations algébriques linéaires avec matrice présentant au plus $m + 1$ éléments non nuls sur chaque ligne.

Avec l'approximation des équations aux dérivées partielles on aboutit également à un système d'équations algébriques discrètes ou ordinaires avec matrice spéciale. Comme le nombre d'inconnues dans un tel système est généralement égal à celui de nœuds du maillage, il arrive qu'on se heurte en pratique à des systèmes d'un ordre

très élevé (à des dizaines et même des centaines de mille d'inconnues). Les autres particularités de ces systèmes sont la raréfaction de la matrice et la structure en bandes, c'est-à-dire une disposition spéciale des éléments non nuls. Ces particularités facilitent, d'une part, la résolution des problèmes mentionnés, mais d'autre part exigent des méthodes de résolution spéciales qui tiendraient compte des caractères spécifiques du problème. Aussi n'est-il pas étonnant que les méthodes classiques de l'algèbre linéaire se révèlent souvent inefficaces dans le cas des équations aux différences et que de plus il n'existe même pas de méthode universelle qui puisse être appliquée efficacement à la résolution de n'importe quelle équation aux différences.

On utilise actuellement deux types de méthodes de résolution des systèmes d'équations algébriques linéaires: 1) les méthodes directes ; 2) les méthodes itératives ou les méthodes d'approximations successives. Habituellement, les méthodes directes sont utilisées à la résolution d'une classe assez étroite d'équations de mailles mais elles permettent d'aboutir à la solution par des calculs relativement peu laborieux. Les méthodes itératives permettent de résoudre des équations plus compliquées et souvent comportent en guise d'étape principale de l'algorithme des méthodes directes de résolution d'équations aux différences spéciales. Le fait que les équations aux différences sont insuffisamment conditionnées implique une mise au point de processus itératifs à convergence rapide avec dégagement du domaine d'efficacité de chaque méthode.

En maintes occasions, par exemple pour des équations linéaires à coefficients constants relativement à la fonction de maille d'un argument, la solution peut être obtenue sous forme fermée. Ces méthodes de résolution des équations de mailles seront l'objet d'étude au § 3 du présent chapitre.

§ 2. Théorie générale des équations aux différences linéaires

1. Propriétés des solutions de l'équation homogène. On étudiera dans ce paragraphe la théorie générale des équations aux différences d'ordre m à coefficients variables

$$a_m(i)\, y(i+m) + \ldots + a_0(i)\, y(i) = f_i,$$

où $a_m(i)$ et $a_0(i)$ sont différents de zéro pour tout i. Passons d'abord à l'étude de l'équation homogène

$$a_m(i)\, y(i+m) + \ldots + a_0(i)\, y(i) = \sum_{k=0}^{m} a_k(i)\, y(i+k) = 0. \qquad (1)$$

Admettons que les coefficients $a_k(i)$, $k = 0, 1, \ldots, m$ possèdent pour toutes les valeurs considérées de i des valeurs finies.

Chaque solution particulière de l'équation (1) est déterminée par les valeurs de la fonction $y(i)$ dans m points variables, mais se disposant successivement i_0, i_0+1, ..., i_0+m-1.

Théorème 1. *Si $v_1(i)$, $v_2(i)$, ..., $v_p(i)$ sont solutions de l'équation* (1), *la fonction*

$$y(i) = c_1v_1(i) + c_2v_2(i) + \ldots + c_pv_p(i), \tag{2}$$

où c_1, c_2, ..., c_p sont des constantes quelconques, est également solution de l'équation (1).

En effet, en vertu de la condition posée par le théorème, on a l'égalité

$$\sum_{k=0}^{m} a_k(i)\, v_l(i+k) = 0, \quad l = 1, 2, \ldots, p. \tag{3}$$

Portons (2) dans (1):

$$\sum_{k=0}^{m} a_k(i)\, y(i+k) = \sum_{k=0}^{m} a_k(i) \sum_{l=1}^{p} c_l v_l(i+k)$$

et modifions l'ordre de sommation dans le second membre de l'égalité. Profitant de (3), il vient

$$\sum_{k=0}^{m} a_k(i)\, y(i+k) = \sum_{l=1}^{p} c_l \sum_{k=0}^{m} a_k(i)\, v_l(i+k) = 0$$

et, par conséquent, la fonction $y(i)$, définie par (2), est également solution de l'équation (1). Le théorème est démontré.

Introduisons la notation $\Delta_i(v_1, \ldots, v_p)$ pour le déterminant

$$\Delta_i(v_1, v_2, \ldots, v_p) = \begin{vmatrix} v_1(i) & v_1(i+1) & \ldots & v_1(i+p-1) \\ v_2(i) & v_2(i+1) & \ldots & v_2(i+p-1) \\ \ldots & \ldots & \ldots & \ldots \\ v_p(i) & v_p(i+1) & \ldots & v_p(i+p-1) \end{vmatrix}.$$

On a le lemme 2.

Lemme 2. *Soient $v_1(i)$, $v_2(i)$, ..., $v_m(i)$ les solutions de l'équation* (1). *Le déterminant $\Delta_i(v_1, \ldots, v_m)$ est soit identiquement nul en i, soit différent de zéro pour toutes les valeurs possibles de i.*

En effet, vu que $v_1(i)$, ..., $v_m(i)$ sont solutions de l'équation (1), les égalités suivantes se vérifient:

$$a_0(i)\, v_1(i) + a_1(i)\, v_1(i+1) + \ldots + a_{m-1}(i)\, v_1(i+m-1) =$$
$$= -a_m(i)\, v_1(i+m),$$

$$a_0(i)\, v_2(i) + a_1(i)\, v_2(i+1) + \ldots + a_{m-1}(i)\, v_2(i+m-1) =$$

$$= -a_m(i)\, v_2(i+m),$$

. . . .

$$a_0(i)\, v_m(i) + a_1(i)\, v_m(i+1) + \ldots + a_{m-1}(i)\, v_m(i+m-1) = $$
$$= -a_m(i)\, v_m(i+m).$$

En résolvant ce système en $a_0(i)$ pour un i fixé à l'aide de la règle de Cramer, il vient

$$a_0(i)\, \Delta_i(v_1, \ldots, v_m) =$$
$$= -a_m(i) \begin{vmatrix} v_1(i+m) & v_1(i+1) & \ldots & v_1(i+m-1) \\ v_2(i+m) & v_2(i+1) & \ldots & v_2(i+m-1) \\ \cdots & \cdots & \cdots & \cdots \\ v_m(i+m) & v_m(i+1) & \ldots & v_m(i+m-1) \end{vmatrix}.$$

Après permutation respective des colonnes du déterminant dans le second membre de l'égalité obtenue on obtient la relation $a_0(i)\, \Delta_i(v_1, \ldots, v_m) = (-1)^m a_m(i)\, \Delta_{i+1}(v_1, \ldots, v_m)$. Comme $a_0(i)$ et $a_m(i)$ ne sont pas nuls pour les valeurs possibles de i, il s'ensuit le lemme énoncé.

Introduisons maintenant la notion de solutions linéairement indépendantes de l'équation (1). Les fonctions de mailles $v_1(i)$, $v_2(i)$, . . ., $v_m(i)$ sont dites *solutions linéairement indépendantes de l'équation* (1) si: 1) elles admettent des valeurs finies et vérifient l'équation (1); 2) la relation

$$c_1v_1(i) + c_2v_2(i) + \ldots + c_mv_m(i) = 0 \qquad (4)$$

pour toutes constantes $c_1, c_2, \ldots, c_m$ simultanément non nulles ne se vérifie pas au moins pour un seul i.

Pour les solutions linéairement indépendantes se vérifie le lemme suivant:

L e m m e 3. *Si* $v_1(i)$, $v_2(i)$, . . ., $v_m(i)$ *sont solutions linéairement indépendantes de l'équation* (1), *le déterminant* $\Delta_i(v_1, \ldots, v_m)$ *est alors non nul pour toutes les valeurs possibles de i. Inversement, si dans les solutions de l'équation* (1) *le déterminant* $\Delta_i(v_1, \ldots, v_m)$ *est différent de zéro au moins pour une des valeurs de i,* $v_1(i)$, . . ., $v_m(i)$ *sont alors des solutions linéairement indépendantes de l'équation* (1).

Vu le lemme 2, le déterminant $\Delta_i(v_1, \ldots, v_m)$ est soit identiquement nul, soit différent de zéro pour tous les i. Soient $v_1(i)$, ..., $v_m(i)$ les solutions linéairement indépendantes de l'équation (1) et admettons que $\Delta_i(v_1, \ldots, v_m) \equiv 0$. Voyons le système d'équations algébriques

$$c_1v_1(i_0) + c_2v_2(i_0) + \ldots + c_mv_m(i_0) = 0.$$
$$c_1v_1(i_0+1) + c_2v_2(i_0+1) + \ldots + c_mv_m(i_0+1) = 0. \qquad (5)$$

. . . .

$$c_1v_1(i_0+m-1) + c_2v_2(i_0+m-1) + \ldots + c_mv_m(i_0 +$$
$$+ m - 1) = 0.$$

Etant donné que le déterminant de ce système $\Delta_{i_0}(v_1, \ldots, v_m)$ est, par hypothèse, nul, il existe une solution de ce système $c_1, c_2, \ldots$ $\ldots, c_m$ différente de zéro. Donc pour les $c_1, c_2, \ldots, c_m$ trouvés on a l'égalité (4) pour $i = i_0, i_0 + 1, \ldots, i_0 + m - 1$. Montrons maintenant que l'égalité (4) a lieu également pour $i = i_0 + m$. A cette fin, en prenant l'équation (1) pour $l = 1, 2, \ldots, m$

$$\sum_{k=0}^{m} a_k(i_0)\, v_l(i_0 + k) = 0,$$

multiplions-la par c_l et sommons les égalités pour $l = 1, 2, \ldots, m$. Tenant compte de l'égalité (5), on obtient

$$0 = a_m(i_0) \sum_{l=1}^{m} c_l v_l(i_0 + m) + \sum_{k=0}^{m-1} a_k(i_0) \sum_{l=1}^{m} c_l v_l(i_0 + k) = \\ = a_m(i_0) \sum_{l=1}^{m} c_l v_l(i_0 + m).$$

On a ainsi démontré la vérité de l'égalité (4) pour $i = i_0 + m$. En raisonnant toujours de la sorte, on obtient que pour les $c_1, c_2, \ldots$ $\ldots, c_m$ trouvés plus haut la relation (4) se vérifie pour tous les $i \geqslant i_0$ possibles. De façon analogue se démontre l'exactitude de (4) pour $i \leqslant i_0$. Donc (4) aux $c_1, c_2, \ldots, c_m$ non nuls se vérifie pour tous les i, ce qui contredit l'indépendance linéaire de $v_1(i), \ldots$ $\ldots, v_m(i)$. Aussi l'hypothèse de ce que le déterminant $\Delta_i(v_1, \ldots$ $\ldots, v_m)$ est identiquement nul en i est erronée.

Passons maintenant à la démonstration de la seconde partie du lemme 3. Supposons que le déterminant $\Delta_i(v_1, \ldots, v_m)$ est, pour un $i = i_0$, différent de zéro. Admettons aussi que $v_1(i), v_2(i), \ldots$ $\ldots, v_m(i)$ est un système de solutions linéairement indépendantes de l'équation (1). Cela signifie qu'il se trouvera des constantes $c_1, c_2, \ldots, c_m$ simultanément non nulles pour lesquelles la relation (4) soit une identité en i. Ecrivons alors (4) pour $i = i_0, i_0 + 1, \ldots$ $\ldots, i_0 + m - 1$ sous forme (5), de plus en raison de l'hypothèse du lemme, le déterminant de ce système $\Delta_{i_0}(v_1, \ldots, v_m)$ est non nul. Tous les $c_1, c_2, \ldots, c_m$ doivent donc être nuls. On aboutit à une contradiction. Le lemme est démontré.

2. Théorèmes sur quelques solutions de l'équation linéaire. Esquissons d'abord la démonstration du théorème de la solution générale de l'équation linéaire homogène (1).

Théorème 2. *Si $v_1(i), v_2(i), \ldots, v_m(i)$ sont solutions linéairement indépendantes de l'équation* (1), *la solution générale de cette équation prend alors la forme*

$$y(i) = c_1 v_1(i) + c_2 v_2(i) + \ldots + c_m v_m(i), \tag{6}$$

où $c_1, c_2, \ldots, c_m$ sont des constantes arbitraires.

En effet, en vertu du théorème 1 la fonction $y(i)$ définie par la formule (6) est solution de l'équation (1). Montrons maintenant que toutes les solutions de l'équation (1) sont contenues dans l'ensemble des fonctions $y(i)$. Soit $u(i)$ une solution quelconque de l'équation (1). Elle est complètement définie si sont données les valeurs initiales aux m points: $u(i_0)$, $u(i_0+1)$, ..., $u(i_0+m-1)$. Choisissons parmi l'ensemble des fonctions de la forme (6) celle qui possède les mêmes valeurs initiales. Il suffit, pour ce faire, de trouver les constantes $c_1, c_2, \ldots, c_m$ pour lesquelles se vérifient les m égalités

$$c_1v_1(i_0) + c_2v_2(i_0) + \ldots + c_mv_m(i_0) = u(i_0),$$
$$c_1v_1(i_0+1) + c_2v_2(i_0+1) + \ldots + c_mv_m(i_0+1) = u(i_0+1),$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$c_1v_1(i_0+m-1) + c_2v_2(i_0+m-1) + \ldots$$
$$\ldots + c_mv_m(i_0+m-1) = u(i_0+m-1).$$

Vu que $v_1(i)$, $v_2(i)$, ..., $v_m(i)$ sont des solutions linéairement indépendantes de (1), en vertu du lemme 3, le déterminant de ce système $\Delta_{i_0}(v_1, \ldots, v_m)$ est donc différent de zéro. En résolvant ce système par rapport à $c_1, c_2, \ldots, c_m$, on obtient la fonction $y(i)$ présentant les mêmes valeurs initiales que $u(i)$. Mais comme les valeurs initiales définissent de façon univoque la solution de l'équation (1), on a $y(i) \equiv u(i)$. Le théorème est démontré.

Examinons maintenant la solution d'une équation inhomogène

$$a_m(i)\,y(i+m) + \ldots + a_0(i)\,y(i) = f(i). \qquad (7)$$

On a le théorème 3.

Théorème 3. *La solution générale de l'équation* (7) *se présente sous forme de somme de sa solution particulière et de la solution générale de l'équation linéaire homogène* (1).

En effet, montrons que toute solution de l'équation (7) peut être représentée sous l'aspect

$$y(i) = \bar{y}(i) + \bar{\bar{y}}(i), \qquad (8)$$

où $\bar{y}(i)$ est une certaine solution de l'équation (7), tandis que $\bar{\bar{y}}(i)$ est la solution générale de l'équation homogène (1). Soit

$$a_m(i)\,\bar{y}(i+m) + \ldots + a_0(i)\,\bar{y}(i) = f(i). \qquad (9)$$

En portant (8) dans (7) et compte tenu de (9), on obtient pour $\bar{\bar{y}}(i)$ l'équation $a_m(i)\,\bar{\bar{y}}(i+m) + \ldots + a_0\bar{\bar{y}}(i) = 0$. Donc $\bar{\bar{y}}(i)$ est la solution générale de l'équation homogène (1). Le théorème est démontré.

C o r o l l a i r e 1. *Il s'ensuit des théorèmes 2 et 3 que la solution générale de l'équation inhomogène (7) a la forme*

$$y(i) = \bar{y}(i) + c_1 v_1(i) + \ldots + c_m v_m(i), \qquad (10)$$

où $\bar{y}(i)$ est la solution particulière de l'équation (7), $v_1(i)$, $v_2(i)$, ..., $v_m(i)$ les solutions linéairement indépendantes de l'équation homogène (1), $c_1, \ldots, c_m$ les constantes arbitraires.

C o r o l l a i r e 2. Utilisant le lemme 3, on peut énoncer le corollaire 1 sous une autre forme: *la solution de l'équation (7) a la forme (10) pour laquelle les solutions particulières $v_1(i), \ldots, v_m(i)$ de l'équation homogène sont telles que $\Delta_i(v_1, \ldots, v_m) \neq 0$ au moins pour une valeur de i.*

C o r o l l a i r e 3. *Si le second membre $f(i)$ de l'équation (7) est la somme de deux fonctions $f(i) = f^{(1)}(i) + f^{(2)}(i)$, la solution particulière de l'équation (7) peut alors être représentée sous forme de $\bar{y}(i) = \bar{y}^{(1)}(i) + \bar{y}^{(2)}(i)$, où $\bar{y}^{(\alpha)}(i)$ est solution particulière de l'équation (7), dont la partie droite est $f^{(\alpha)}(i)$, avec $\alpha = 1, 2$.*

3. Méthode de variation des constantes. Les théorèmes démontrés plus haut esquissent la structure de la solution générale de l'équation aux différences linéaire inhomogène (7). Abordons maintenant les questions suivantes: 1) comment construire les solutions linéairement indépendantes d'une équation homogène; 2) comment obtenir la solution particulière d'une équation inhomogène; 3) comment, en utilisant la solution générale de l'équation inhomogène, peut-on aboutir à une solution unique de l'équation (7) qui satisfait aux conditions complémentaires.

Etudions d'abord un procédé réalisable de construction de solutions linéairement indépendantes d'une équation homogène. Etant donné que la solution particulière d'une équation linéaire d'ordre m se définit complètement par la fixation des valeurs initiales en m points, par exemple, $i = i_0, i_0 + 1, \ldots, i_0 + m - 1$, on peut, en partant du lemme 3, construire les solutions cherchées de la façon suivante. Soit A la matrice non dégénérée

$$A = \begin{Vmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mm} \end{Vmatrix}.$$

Construisons m solutions de l'équation (1) $v_1(i), v_2(i), \ldots, v_m(i)$ définies par les valeurs initiales

$$v_l(i_0 + k - 1) = a_{lk}, \quad l, k = 1, 2, \ldots, m. \qquad (11)$$

Dans ce cas $\Delta_{i_0}(v_1, \ldots, v_m) = \det A \neq 0$. Donc le problème de construction des fonctions cherchées $v_1(i), \ldots, v_m(i)$ est résolu.

Examinons maintenant la question d'isolation de la solution unique de la famille de solutions (10). Il s'ensuit de (10) qu'il faut pour cela définir m conditions exactement sur la fonction $y(i)$, sur la base desquelles on définit les constantes $c_1, c_2, \ldots, c_m$.

Dans le cas du problème de Cauchy, c'est-à-dire quand sont données les conditions initiales $y(i_0) = b_1$, $y(i_0+1) = b_2, \ldots$ $\ldots, y(i_0+m-1) = b_m$, la détermination des constantes c_1, $c_2, \ldots, c_m$ s'effectue sans peine. De (10) on obtient un système d'équations algébriques linéaires relativement à $c_1, c_2, \ldots, c_m$:

$$\begin{gathered} v_1(i_0)\,c_1 + v_2(i_0)\,c_2 + \ldots + v_m(i_0)\,c_m = b_1 - \bar{y}(i_0), \\ v_1(i_0+1)\,c_1 + v_2(i_0+1)\,c_2 + \ldots + v_m(i_0+1)\,c_m = \\ = b_2 - \bar{y}(i_0+1), \qquad (12) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ v_1(i_0+m-1)\,c_1 + v_2(i_0+m-1)\,c_2 + \ldots \\ \ldots + v_m(i_0+m-1)\,c_m = b_m - \bar{y}(i_0+m-1). \end{gathered}$$

Comme le déterminant de ce système $\Delta_{i_0}(v_1, \ldots, v_m)$ est différent de zéro, ce système ne possède qu'une seule solution $c_1, c_2, \ldots, c_m$ qui détermine complètement la solution unique de l'équation inhomogène (7).

Au cas du problème aux limites pour lequel les m conditions supplémentaires imposées à $y(i)$ sont données de façon quelconque par rapport aux points successifs, on aboutit de nouveau au système d'équations algébriques linéaires en $c_1, c_2, \ldots, c_m$. Mais dans ce cas la solution de ce système ne s'obtiendra qu'à la condition de nouvelles hypothèses relativement aux coefficients de l'équation aux différences.

Voyons maintenant la question de la solution des équations (12). Puisqu'en vertu de (11) la matrice du système (12) est A^T, en choisissant en guise de A la matrice unité on obtient la solution du système (12) sous forme explicite: $c_l = b_l - \bar{y}(i_0 + l - 1)$, $l = 1, 2, \ldots, m$. Il est apparemment rationnel de choisir une telle solution parmi les solutions particulières de l'équation inhomogène (7) pour laquelle $\bar{y}(i_0) = \bar{y}(i_0+1) = \ldots = \bar{y}(i_0+m-1) = 0$. On aura alors $c_l = b_l$, $l = 1, 2, \ldots, m$.

A ce choix de la matrice A correspondent les valeurs initiales de $v_1(i), \ldots, v_m(i)$ suivantes:

$$v_l(i_0 + l - 1) = 1, \quad v_l(i_0 + k - 1) = 0, \quad k = 1, 2, \ldots, m, \\ k \neq l, \quad l = 1, 2, \ldots, m.$$

Occupons-nous maintenant de la recherche des solutions particulières de l'équation inhomogène si sont connues m solutions linéairement indépendantes de l'équation homogène. Exposons le procédé

de recherche de la solution particulière par *variation des constantes* dans la solution générale de l'équation homogène.

On a montré auparavant que la solution générale de l'équation homogène (1) a l'aspect suivant: $\bar{\bar{y}}(i) = c_1 v_1(i) + \ldots + c_m v_m(i)$, où $v_1(i), \ldots, v_m(i)$ est la solution linéairement indépendante de l'équation (1) et $c_1, c_2, \ldots, c_m$ les constantes arbitraires. Admettons maintenant que $c_1, c_2, \ldots, c_m$ sont des fonctions de i et posons le problème de leur choix de manière que la fonction

$$\bar{y}(i) = c_1(i)\, v_1(i) + \ldots + c_m(i)\, v_m(i) \tag{13}$$

soit une solution particulière de l'équation inhomogène (7). Notons que chaque fonction $c_l(i)$ se définit à la précision jusqu'à la constante près, vu que $v_l(i)$ est la solution de l'équation homogène:

$$a_m(i)\, v_l(i+m) + \ldots + a_0(i)\, v_l(i) = 0, \quad l = 1, 2, \ldots, m. \tag{14}$$

Introduisons la notation suivante:

$$d_k(i) = \sum_{l=1}^{m} [c_l(i+k) - c_l(i)]\, v_l(i+k), \quad k = 0, 1, \ldots, m.$$

Portant (13) dans (7), exécutant les transformations identiques dans l'expression obtenue et tenant compte de (14), il vient

$$f(i) = \sum_{k=0}^{m} a_k(i)\, \bar{y}(i+k) = \sum_{k=0}^{m} a_k(i) \sum_{l=1}^{m} c_l(i+k)\, v_l(i+k) =$$

$$= \sum_{k=0}^{m} a_k(i)\, d_k(i) + \sum_{k=0}^{m} a_k(i) \sum_{l=1}^{m} c_l(i)\, v_l(i+k) =$$

$$= \sum_{k=0}^{m} a_k(i)\, d_k(i) + \sum_{l=1}^{m} c_l(i) \left[\sum_{k=0}^{m} a_k(i)\, v_l(i+k)\right] =$$

$$= \sum_{k=0}^{m} a_k(i)\, d_k(i) = \sum_{k=1}^{m} a_k(i)\, d_k(i),$$

car $d_0(i) \equiv 0$. La relation obtenue se vérifie pour tous les i si l'on pose

$$d_k(i) = 0, \quad k = 1, 2, \ldots, m-1, \quad d_m(i) = f(i)/a_m(i). \tag{15}$$

Bref, le problème de la construction des fonctions $c_1(i), c_2(i), \ldots, c_m(i)$ se réduit à leur détermination à partir des conditions (15) qui doivent se réaliser identiquement en i.

Transformons le système d'équations (15). Posons $b_l(i) = c_l(i+1) - c_l(i)$, $l = 1, 2, \ldots, m$. De la définition de $d_k(i)$ on obtient pour $k = 1, 2, \ldots, m$:

$$d_k(i) - d_{k-1}(i+1) = \sum_{l=1}^{m} [c_l(i+k) - c_l(i)]\, v_l(i+k) -$$

$$- \sum_{l=1}^{m} [c_l(i+k) - c_l(i+1)]\, v_l(i+k) = \sum_{l=1}^{m} b_l(i)\, v_l(i+k).$$

En y portant (15) et compte tenu de l'égalité $d_0(i) = 0$, on obtient le système d'équations algébriques linéaires relativement à $b_l(i)$ pour un i fixé:

$$\begin{aligned} &b_1(i)\, v_1(i+1) + b_2(i)\, v_2(i+1) + \ldots + b_m(i)\, v_m(i+1) = 0, \\ &b_1(i)\, v_1(i+2) + b_2(i)\, v_2(i+2) + \ldots + b_m(i)\, v_m(i+2) = 0, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \qquad (16) \\ &b_1(i)\, v_1(i+m) + b_2(i)\, v_2(i+m) + \ldots + b_m(i)\, v_m(i+m) = \frac{f(i)}{a_m(i)}. \end{aligned}$$

Le déterminant du système (16) est égal à $\Delta_{i+1}(v_1, v_2, \ldots, v_m)$ et est différent de zéro en raison des indépendances linéaires $v_1, v_2, \ldots, v_m$. Aussi le système (16) n'a-t-il qu'une solution unique

$$b_l(i) = c_l(i+1) - c_l(i) = (-1)^{m+l} \frac{f(i)}{a_m(i)} \frac{\mathcal{D}_l(i)}{\mathcal{D}(i)}, \quad l = 1, \ldots, m, \qquad (17)$$

où $\mathcal{D}(i) = \Delta_{i+1}(v_1, v_2, \ldots, v_m)$, et

$$\mathcal{D}_l(i) = \\ = \begin{vmatrix} v_1(i+1) & \ldots & v_{l-1}(i+1) & v_{l+1}(i+1) & \ldots & v_m(i+1) \\ v_1(i+2) & \ldots & v_{l-1}(i+2) & v_{l+1}(i+2) & \ldots & v_m(i+2) \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ v_1(i+m-1) & \ldots & v_{l-1}(i+m-1) & v_{l+1}(i+m-1) & \ldots & v_m(i+m-1) \end{vmatrix},$$

autrement dit, $\mathcal{D}_l(i)$ s'obtient à partir du déterminant $\mathcal{D}(i)$ par élimination de la l-ième colonne et de la dernière ligne.

Les égalités (17) sont des équations aux différences du premier ordre relativement aux fonctions $c_l(i)$, $l = 1, 2, \ldots, m$. Vu que $c_l(i)$ peut être défini à la précision de la constante près, on obtient à partir de (17) la représentation explicite de $c_l(i)$:

$$c_l(i) = \sum_{j=i_0}^{i-1} (-1)^{m+l} \frac{f(j)\, \mathcal{D}_l(j)}{a_m(j)\, \mathcal{D}(j)}, \quad l = 1, 2, \ldots, m.$$

En portant cette expression dans (13) et en changeant l'ordre de la sommation dans la représentation obtenue, on obtient pour la solution particulière $\bar{y}(i)$ de l'équation inhomogène (7) la formule suivante:

$$\begin{aligned} \bar{y}(i) &= \sum_{l=1}^{m} c_l(i)\, v_l(i) = \\ &= \sum_{j=i_0}^{i-1} \left[f(j) \sum_{l=1}^{m} (-1)^{m+l} \mathcal{D}_l(j)\, v_l(i) \right] / (\mathcal{D}(j)\, a_m(j)) = \\ &= \sum_{j=i_0}^{i-1} G(i, j)\, f(j), \end{aligned}$$

où

$$G(i, j)=\frac{1}{\mathscr{D}(j)\,a_m(j)}\sum_{k=1}^{m}(-1)^{m+k}\,\mathscr{D}_k(j)\,v_k(i). \tag{18}$$

Notons que la somme figurant dans (18) se calcule aisément

$$\sum_{k=1}^{m}(-1)^{m+k}\,\mathscr{D}_k(j)\,v_k(i)=$$

$$=\begin{vmatrix} v_1(j+1) & v_2(j+1) & \dots & v_m(j+1)\\ v_1(j+2) & v_2(j+2) & \dots & v_m(j+2)\\ \dots & \dots & \dots & \dots\\ v_1(j+m-1) & v_2(j+m-1) & \dots & v_m(j+m-1)\\ v_1(i) & v_2(i) & \dots & v_m(i)\end{vmatrix}.$$

Cette somme est nulle pour $j=i-1,\ i-2,\ \dots,\ i-m+1$. Donc la solution particulière de l'équation (7) a la représentation suivante

$$\bar{y}(i)=\sum_{j=i_0}^{i-m}\frac{\begin{vmatrix} v_1(j+1) & \dots & v_m(j+1)\\ \dots & \dots & \dots\\ v_1(j+m-1) & \dots & v_m(j+m-1)\\ v_1(i) & \dots & v_m(i)\end{vmatrix}}{\begin{vmatrix} v_1(j+1) & \dots & v_1(j+m)\\ \dots & \dots & \dots\\ v_m(j+1) & \dots & v_m(j+m)\end{vmatrix}}\,\frac{f(j)}{a_m(j)}, \tag{19}$$

où i_0 est quelconque, tandis que pour $i=i_0,\ i_0+1,\ \dots,\ i_0+m-1$, on a $\bar{y}(i)=0$.

Pour une équation du premier ordre ($m=1$) la formule (19) prend la forme suivante:

$$\bar{y}(i)=\sum_{j=i_0}^{i-1}\frac{v_1(i)}{v_1(j+1)}\cdot\frac{f(j)}{a_1(j)},\qquad \bar{y}(i_0)=0. \tag{20}$$

4. Exemples. Examinons quelques exemples illustrant l'application de la théorie générale. Supposons qu'il s'agit de trouver la solution générale de l'équation du premier ordre

$$y(i+1)-e^{2i}y(i)=6i^2e^{i^2+i}. \tag{21}$$

Cherchons d'abord la solution de l'équation homogène

$$y(i+1)-e^{2i}y(i)=0. \tag{22}$$

A partir de (22) on obtient successivement

$$y(i+1)=e^{2i}y(i) = e^{2i}e^{2(i-1)}y(i-1) = \ldots = e^{2\sum\limits_{k=1}^{i} k}\, y(1) = e^{i(i+1)}y(1).$$

En posant ici $y(1) = 1$, on trouve la solution particulière $v_1(i)$ de l'équation homogène (22) sous la forme $v_1(i) = e^{i(i-1)}$. La solution générale de l'équation homogène a donc la forme $\bar{\bar{y}}(i) = ce^{i(i-1)}$, où c est une constante arbitraire.

Construisons maintenant la solution particulière de l'équation inhomogène (21) sur la base de la formule (20). De (20) il vient

$$\bar{y}(i) = \sum_{k=i_0}^{i-1} \frac{e^{i(i-1)}}{e^{k(k+1)}} \cdot \frac{6k^2e^{k^2+k}}{1} = 6e^{i(i-1)} \sum_{k=i_0}^{i-1} k^2.$$

Vu que i_0 peut être choisi quelconque, en posant ici $i_0 = 1$, on obtient $\bar{y}(i) = i(i-1)(2i-1)e^{i(i-1)}$. Ensuite, en vertu du théorème 3, la solution générale de l'équation (21) s'écrit sous la forme

$$y(i) = \bar{y}(i) + \bar{\bar{y}}(i) = [c + i(i-1)(2i-1)]\, e^{i(i-1)},$$

où c est une constante arbitraire. Le problème est résolu.

Cherchons maintenant la solution générale de l'équation du second ordre

$$a_2(i)\, y(i+2) + a_1(i)\, y(i+1) + a_0(i)\, y(i) = f(i), \tag{23}$$

où $i = 0, 1, 2, \ldots$,

$$\begin{aligned} &a_2(i) = i^2 - i + 1,\quad a_0(i) = a_2(i+1) = i^2 + i + 1,\\ &a_1(i) = -a_0(i) - a_2(i) = -2(i^2+1),\\ &f(i) = 2^i(i^2 - 3i + 1) = 2^i[2a_2(i) - a_0(i)]. \end{aligned} \tag{24}$$

Vu que les coefficients $a_2(i)$ et $a_0(i)$ sont différents de zéro, pour trouver la solution générale de l'équation (23) on peut appliquer la théorie générale.

Construisons d'abord les solutions linéairement indépendantes de l'équation homogène. Sur la base de (24) elle peut être écrite sous la forme:

$$a_2(i)\, y(i+2) - [a_2(i) + a_2(i+1)]\, y(i+1) + a_2(i+1)\, y(i) = 0$$

ou

$$a_2(i)\, [y(i+2) - y(i+1)] - a_2(i+1)\, [y(i+1) - y(i)] = 0. \tag{25}$$

Les solutions particulières $v_1(i)$ et $v_2(i)$ de l'équation homogène (25) seront isolées par les conditions suivantes: $v_1(0) = v_1(1) = 1$, $v_2(0) = 0$, $v_2(1) = 3$. Puisque le déterminant

$$\Delta_0(v_1, v_2) = \begin{vmatrix} v_1(0) & v_1(1) \\ v_2(0) & v_2(1) \end{vmatrix} = 3 \neq 0,$$

en vertu du lemme 3 les fonctions $v_1(i)$ et $v_2(i)$ seront des solutions linéairement indépendantes de l'équation (25).

Cherchons la forme explicite de $v_1(i)$ et $v_2(i)$. Il s'ensuit directement de (25) que $v_1(i) \equiv 1$. Construisons $v_2(i)$. De (25) on obtient successivement

$$y(i+2) - y(i+1) = \frac{a_2(i+1)}{a_2(i)}[y(i+1) - y(i)] =$$

$$= \frac{a_2(i+1)}{a_2(i-1)}[y(i) - y(i-1)] = \ldots = \frac{a_2(i+1)}{a_2(0)}[y(1) - y(0)].$$

Compte tenu des valeurs initiales de $v_2(i)$, on obtient de ce qui précède

$$v_2(i+1) - v_2(i) = 3a_2(i) = 3(i^2 - i + 1). \tag{26}$$

En sommant les parties gauche et droite de (26) en i de zéro à $k - 1$, il vient

$$v_2(k) = v_2(0) + 3\sum_{i=0}^{k-1}(i^2 - i + 1) = k(k^2 - 3k + 5).$$

Bref, les solutions particulières de l'équation homogène (25) sont trouvées

$$v_1(k) \equiv 1, \quad v_2(k) = k(k^2 - 3k + 5), \tag{27}$$

et la solution générale de (25) prend la forme $\overset{=}{y}(k) = c_1 + c_2k(k^2 - 3k + 5)$.

Construisons maintenant la solution particulière de l'équation inhomogène (23). Portant (24) et (27) dans la formule (19), il vient

$$\bar{y}(i) = \sum_{k=0}^{i-2} \frac{v_2(i) - v_2(k+1)}{v_2(k+2) - v_2(k+1)} \cdot \frac{f(k)}{a_2(k)} =$$

$$= \sum_{k=0}^{i-2} \frac{v_2(i) - v_2(k+1)}{3a_2(k+1)\,a_2(k)}[2^{k+1}a_2(k) - 2^k a_2(k+1)] =$$

$$= \frac{1}{3}\sum_{k=0}^{i-2}[v_2(i) - v_2(k+1)]\left[\frac{2^{k+1}}{a_2(k+1)} - \frac{2^k}{a_2(k)}\right]. \tag{28}$$

On a utilisé ici l'égalité (26).

Calculons l'expression obtenue. En notant

$$v(k)=v_2(i)-v_2(k+1),\quad u(k)=\frac{2^k}{a_2(k)},$$

écrivons (28) de la façon suivante:

$$\bar{y}(i)=\frac{1}{3}\sum_{k=0}^{i-2}[u(k+1)-u(k)]\,v(k).$$

Appliquons maintenant la formule de sommation par parties (voir (8) § 1) pour le cas d'un maillage régulier de pas $h=1$. On a

$$\bar{y}(i)=-\frac{1}{3}\sum_{k=0}^{i-1}u(k)\,[v(k)-v(k-1)]+$$
$$+\frac{1}{3}[u(i-1)\,v(i-1)-u(0)\,v(-1)].$$

Vu qu'en raison de (26), de la condition $v_2(0)=0$ et de la définition des fonctions $v(k)$ et $u(k)$ on a

$$v(k)-v(k-1)=v_2(k)-v_2(k+1)=-3a_2(k),$$
$$v(i-1)=v_2(i)-v_2(i)=0,$$
$$v(-1)=v_2(i)-v_2(0)=v_2(i),$$

il s'ensuit que

$$\bar{y}(i)=\sum_{k=0}^{i-1}2^k-\frac{1}{3}v_2(i)=2^i-1-\frac{1}{3}i\,(i^2-3i+5).$$

On a donc trouvé la solution particulière de (23). En vertu du théorème 3 la solution générale de l'équation inhomogène d'ordre deux (23) a la forme

$$y(i)=\bar{y}(i)+\bar{\bar{y}}(i)=2^i-1-\frac{1}{3}i\,(i^2-3i+5)+c_1+c_2i\,(i^2-3i+5)=$$
$$=\bar{c}_1+2^i+\bar{c}_2i\,(i^2-3i+5),$$

où $\bar{c}_1=c_1-1,\ \bar{c}_2=c_2-\frac{1}{3}$ sont des constantes arbitraires. Le problème est résolu.

§ 3. Solution des équations linéaires à coefficients constants

1. Equation caractéristique. Cas de racines simples. Etudions maintenant une classe importante d'équations aux différences, les équations linéaires à coefficients constants. Pour les équations de

cette classe le problème de la recherche des solutions linéairement indépendantes d'équations homogènes adéquates se résout de façon assez simple. Or, comme il a été montré plus haut, c'est à quoi aboutit le problème de recherche de la solution de l'équation aux différences inhomogène.

Recherchons les solutions linéairement indépendantes des équations linéaires homogènes à coefficients constants d'ordre m

$$a_m y(i+m) + a_{m-1} y(i+m-1) + \ldots + a_0 y(i) = 0. \qquad (1)$$

Cherchons les solutions particulières (1) sous forme $v(i) = q^i$, où le nombre q doit être défini. En substituant $v(i)$ à $y(i)$ dans (1), on obtient l'équation

$$q^i (a_m q^m + a_{m-1} q^{m-1} + \ldots + a_1 q + a_0) = 0.$$

Comme on ne recherche pas la solution identiquement nulle de (1), en simplifiant par q^i, on obtient de la dernière expression l'équation pour q:

$$a_m q^m + a_{m-1} q^{m-1} + \ldots + a_1 q + a_0 = 0. \qquad (2)$$

L'équation (2) s'appelle *équation caractéristique* de (1). Les racines de l'équation (2) $q_1, q_2, \ldots, q_m$ peuvent être soit simples soit multiples. Examinons séparément chaque cas éventuel.

Soient des racines simples. Montrons que les fonctions

$$v_1(i) = q_1^i, \quad v_2(i) = q_2^i, \ldots, v_m(i) = q_m^i \qquad (3)$$

sont des solutions linéairement indépendantes de (1).

En effet, en vertu du lemme 3 il suffit de montrer qu'au moins pour un i le déterminant $\Delta_i(v_1, v_2, \ldots, v_m) \neq 0$. En posant $i = 0$, il vient

$$\Delta_0(v_1, \ldots, v_m) = \begin{vmatrix} 1 & q_1 & q_1^2 & \ldots & q_1^{m-1} \\ 1 & q_2 & q_2^2 & \ldots & q_2^{m-1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & q_m & q_m^2 & \ldots & q_m^{m-1} \end{vmatrix} = \begin{vmatrix} 1 & 1 & \ldots & 1 \\ q_1 & q_2 & \ldots & q_m \\ q_1^2 & q_2^2 & \ldots & q_m^2 \\ \ldots & \ldots & \ldots & \ldots \\ q_1^{m-1} & q_2^{m-1} & \ldots & q_m^{m-1} \end{vmatrix}$$

et, par conséquent, $\Delta_0(v_1, \ldots, v_m)$ est le déterminant de Vandermonde. Il est différent de zéro, vu que tous les q_h sont différents. Donc les fonctions (3) sont en fait des solutions de (1) linéairement indépendantes et, c'est pourquoi, la solution générale de l'équation homogène (1) peut être écrite sous la forme

$$y(i) = c_1 q_1^i + c_2 q_2^i + \ldots + c_m q_m^i, \qquad (4)$$

où $c_1, c_2, \ldots, c_m$ sont des constantes arbitraires.

Si les racines $q_1, q_2, \ldots, q_m$ sont réelles, la solution réelle $y(i)$ s'explicite par le choix des constantes $c_1, c_2, \ldots, c_m$ sous forme de

nombres réels. Etudions maintenant le problème de l'explicitation de la solution réelle au cas où, parmi les racines, il y a des racines complexes.

Soit $q_n = \rho\,(\cos\varphi + i^* \sin\varphi)$, $(i^* = \sqrt{-1})$ — la racine complexe de l'équation caractéristique (2). Il existe alors une racine $q_s = \rho\,(\cos\varphi - i^* \sin\varphi)$ conjuguée à q_n. Etudions la partie de la solution générale (4) constituée par la combinaison linéaire de q_n^i et q_s^i:

$$y\,(i) = c_n q_n^i + c_s q_s^i = \rho^i\,[(c_n + c_s)\cos i\varphi + i^*\,(c_n - c_s)\sin i\varphi].$$

La fonction $y\,(i)$ aura des valeurs réelles si les constantes c_n et c_s seront des nombres conjugués complexes. En posant $c_n = 0{,}5\,(\bar{c}_n - i^*\bar{c}_s)$, $c_s = 0{,}5\,(\bar{c}_n + i^*\bar{c}_s)$, où $\bar{c}_n$ et $\bar{c}_s$ sont des nombres réels quelconques, on obtient $y\,(i) = \rho^i\,(\bar{c}_n \cos i\varphi + \bar{c}_s \sin i\varphi)$.

2. **Cas de racines multiples.** Soit maintenant l'équation caractéristique (2) aux racines q_1 de multiplicité n_1, q_2 de multiplicité n_2, etc., autrement dit, $q_1, q_2, \ldots, q_s$ sont des racines différentes de multiplicité respectivement $n_1, n_2, \ldots, n_s$, $n_1 + n_2 + \ldots + n_s = m$. Construisons les solutions linéairement indépendantes de l'équation homogène (1). Recourrons au

Lemme 4. *Si q_l est la racine de l'équation caractéristique* (2) *possédant la multiplicité n_l, alors les égalités*

$$\sum_{k=0}^{m} a_k k^p q_l^k = 0, \qquad p = 0, 1, \ldots, n_l - 1 \tag{5}$$

se vérifient.

En effet, q_l étant la racine de l'équation (2) de multiplicité n_l, on a les égalités

$$\sum_{k=0}^{m} a_k q_l^k = 0, \tag{6}$$

$$\sum_{k=0}^{m} k\,(k-1)\ldots(k-s+1)\,a_k q_l^k = 0, \qquad s = 1, 2, \ldots, n_l - 1, \tag{7}$$

tirées à partir de (2) par s dérivations et multiplication complémentaire du résultat par q_l^s. Montrons que l'égalité (5) est équivalente à (6), (7). Pour cela, il suffit de démontrer l'équivalence de (7) et (5) pour $p \geqslant 1$.

Vu que $P_s\,(k) = k\,(k-1)\ldots(k-s+1)$ est un polynôme du s-ième degré, en multipliant (5) par le coefficient correspondant du polynôme $P_s\,(k)$ pour $p = 1, 2, \ldots, s$ et en additionnant les égalités ainsi obtenues, on aboutit à la relation (7).

Montrons maintenant que de (7) s'ensuivent les égalités (5) pour $p = 1, 2, \ldots, n_l - 1$. Profitons du développement en k^p:

$$k^p = \sum_{s=1}^{p} k(k-1)\ldots(k-s+1)\alpha_s, \quad 1 \leqslant p \leqslant k, \tag{8}$$

où $\alpha_s = \alpha_s(p)$ sera précisé plus loin. Multiplions la s-ième égalité de (7) par α_s et sommons en s de 1 à p. En vertu de (8), il vient

$$0 = \sum_{s=1}^{p} \alpha_s \Big(\sum_{k=0}^{m} k(k-1)\ldots(k-s+1)\, a_k q_l^k \Big) =$$

$$= \sum_{k=0}^{m} a_k q_l^k \Big(\sum_{s=1}^{p} k(k-1)\ldots(k-s+1)\alpha_s \Big) = \sum_{k=0}^{m} a_k k^p q_l^k.$$

Il nous reste à argumenter le développement (8). Notons qu'à gauche et à droite dans (8) figurent des polynômes du p-ième degré en k. Si l'on pose $\alpha_p = 1$, les coefficients de degré supérieur de k seront égaux dans (8) à gauche comme à droite, tandis que les coefficients de degré inférieur de k seront nuls. Cherchons $\alpha_1, \alpha_2, \ldots$ $\ldots, \alpha_{p-1}$ en égalant les valeurs des polynômes en $p - 1$ points différents, par exemple, en posant $k = 1, 2, \ldots, p - 1$. Pour $k = 1$, on a $\alpha_1 = 1$. Pour $k = n$, $2 \leqslant n \leqslant p - 1$, il vient

$$n^p = \sum_{s=1}^{p} n(n-1)\ldots(n-s+1)\alpha_s = \sum_{s=1}^{n} n(n-1)\ldots(n-s+1)\alpha_s =$$

$$= n!\,\alpha_n + n! \sum_{s=1}^{n-1} \frac{\alpha_s}{(n-s)!}.$$

Cette expression permet de trouver α_n si $\alpha_1, \alpha_2, \ldots, \alpha_{n-1}$ sont déjà déterminés. On obtient ainsi la formule de récurrence suivante permettant de trouver les coefficients α_n:

$$\alpha_n = \frac{n^p}{n!} - \sum_{s=1}^{n-1} \frac{\alpha_s}{(n-s)!}, \quad n = 2, 3, \ldots, p-1, \quad \alpha_1 = 1.$$

Le lemme est démontré.

En utilisant le lemme 4, on trouve m solutions particulières de l'équation homogène (1). L'équation

$$(j+k)^n = \sum_{p=0}^{n} C_n^p k^p j^{n-p}, \quad C_n^p = \frac{n!}{p!\,(n-p)!}$$

étant vérifiée, en multipliant (5) par $C_n^p\, j^{n-p} q_l^j$ et en sommant en p de zéro à $n \leqslant n_l - 1$, on obtient pour tout j les égalités

$$\sum_{k=0}^{m} a_k (j+k)^n q_l^{k+j} = 0, \quad n = 0, 1, \ldots, n_l - 1.$$

En les utilisant, on trouve aisément que les solutions particulières de l'équation homogène (1) sont des fonctions de mailles

$$v_{n_1+n_2+\ldots+n_{l-1}+n+1}(j)=j^n q_l^j, \quad 0\leqslant n\leqslant n_l-1, \quad l=1,2,\ldots,s, \tag{9}$$

c'est-à-dire que si q_l est la racine de l'équation caractéristique de multiplicité n_l, les fonctions

$$q_l^j, jq_l^j, \ldots, j^{n_l-1}q_l^j, \quad l=1,2,\ldots,s$$

sont alors solutions de l'équation (1).

Il reste à montrer que les fonctions $v_1(j), \ldots, v_m(j)$, définies dans (9), sont des solutions linéairement indépendantes. A cette fin calculons le déterminant $\Delta_0(v_1, \ldots, v_m)$ qui, dans le cas considéré, a l'aspect

$$\Delta_0(v_1,\ldots,v_m)=\begin{vmatrix} 1 & q_1 & q_1^2 & \ldots & q_1^k & \ldots & q_1^{m-1} \\ 0 & q_1 & 2q_1^2 & \ldots & kq_1^k & \ldots & (m-1)\,q_1^{m-1} \\ 0 & q_1 & 2^2q_1^2 & \ldots & k^2q_1^k & \ldots & (m-1)^2\,q_1^{m-1} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & q_2 & q_2^2 & \ldots & q_2^k & \ldots & q_2^{m-1} \\ 0 & q_2 & 2q_2^2 & \ldots & kq_2^k & \ldots & (m-1)\,q_2^{m-1} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & q_s & 2^{n_s-1}q_s^2 & \ldots & k^{n_s-1}q_s^k & \ldots & (m-1)^{n_s-1}\,q_s^{m-1} \end{vmatrix}.$$

Il peut être obtenu directement à partir du déterminant de Vandermonde

$$W(x_1, x_2, \ldots, x_m)=\begin{vmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^{m-1} \\ 1 & x_2 & x_2^2 & \ldots & x_2^{m-1} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{m-1} & x_{m-1}^2 & \ldots & x_{m-1}^{m-1} \\ 1 & x_m & x_m^2 & \ldots & x_m^{m-1} \end{vmatrix}=\prod_{i=1}^{m-1}\prod_{j=i+1}^{m}(x_j-x_i)$$

de la façon suivante. Prenons à partir de W la dérivée première en x_2 et multiplions-la par x_2. Désignons le résultat par $W_2=$ $=x_2\frac{\partial W}{\partial x_2}$. Ensuite, calculons

$$W_3=x_3\frac{\partial}{\partial x_3}\left(x_3\frac{\partial W_2}{\partial x_3}\right), \quad W_4=x_4\frac{\partial}{\partial x_4}\left(x_4\frac{\partial}{\partial x_4}\left(x_4\frac{\partial W_3}{\partial x_4}\right)\right), \ldots, \text{etc.},$$

tant qu'on n'obtienne W_{n_1}. Ensuite, calculons $W_{n_1+2}=x_{n_1+2}\times$ $\times\frac{\partial W_{n_1}}{\partial x_{n_1+2}}$ et continuons les opérations de dérivation en calculant $W_{n_1+3}=x_{n_1+3}\frac{\partial}{\partial x_{n_1+3}}\left(x_{n_1+3}\frac{\partial W_{n_1+2}}{\partial x_{n_1+3}}\right)$, tant qu'on n'obtienne $W_{n_1+n_2}$, etc. On obtient finalement $W_m=W_m(x_1, x_2, \ldots, x_m)$.

Posons ici $x_1 = x_2 = \ldots = x_{n_1} = q_1$, $x_{n_1+1} = x_{n_1+2} = \ldots = x_{n_1+n_2} = q_2$, etc. On se convainc sans peine que $\Delta_0 (v_1, v_2, \ldots, v_m) = W_m$, tandis que les calculs élémentaires donnent

$$W_m = \prod_{k=1}^{s} \prod_{m=1}^{n_k-1} m!\, q_k^m \prod_{i=1}^{s-1} \prod_{j=i+1}^{s} (q_j - q_i)^{n_i n_j}.$$

Il s'ensuit que $\Delta_0 (v_1, \ldots, v_m) \neq 0$, vu que $q_j \neq q_i$ pour $j \neq i$, et, par suite, les fonctions $v_1 (j)$, $v_2 (j)$, ..., $v_m (j)$, construites plus haut, sont des solutions linéairement indépendantes de l'équation homogène (1). Dans ce cas la solution générale de l'équation (1) s'écrit sous la forme

$$y(j) = \sum_{l=1}^{s} \sum_{n=0}^{n_p-1} c_n^{(l)} j^n q_l^j,$$

où $c_n^{(l)}$ sont des constantes arbitraires.

3. Exemples. Examinons les plus simples exemples de recherche de la solution générale d'une équation aux différences homogène à coefficients constants.

1. Il s'agit de trouver la solution générale de l'équation

$$y(i+2) - y(i+1) - 2y(i) = 0. \qquad (10)$$

Composons l'équation caractéristique $q^2 - q - 2 = 0$ et cherchons ses racines $q_1 = 2$, $q_2 = -1$. Les racines étant simples, la solution générale de l'équation (10) prend l'aspect

$$y(i) = c_1 2^i + c_2 (-1)^i.$$

2. Trouver la solution générale de l'équation d'ordre 4

$$y(j+4) - 2y(j+3) + 3y(j+2) + 2y(j+1) - 4y(j) = 0. \qquad (11)$$

L'équation caractéristique $q^4 - 2q^3 + 3q^2 + 2q - 4 = 0$ possède deux racines réelles $q_1 = 1$, $q_2 = -1$ et deux racines complexes conjuguées $q_3 = 2\left(\cos\frac{\pi}{3} + i \sin\frac{\pi}{3}\right)$ et $q_4 = 2\left(\cos\frac{\pi}{3} - i \sin\frac{\pi}{3}\right)$, $i = \sqrt{-1}$. La solution générale de l'équation (11), prenant des valeurs réelles, a donc l'aspect

$$y(j) = c_1 + c_2 (-1)^j + 2^j \left(c_3 \cos\frac{\pi}{3} j + c_4 \sin\frac{\pi}{3} j\right).$$

3. Trouver la solution générale de l'équation d'ordre 4

$$y(j+4) - 7y(j+3) + 18y(j+2) - 20y(j+1) + 8y(j) = 0. \qquad (12)$$

L'équation caractéristique

$$q^4 - 7q^3 + 18q^2 - 20q + 8 = (q-2)^3(q-1) = 0$$

possède une racine $q_1 = 2$ de multiplicité 3 et une racine $q_2 = 1$ de multiplicité 1. Donc la solution générale de (12) a la forme

$$y(j) = c_1 + 2^j(c_2 + c_3 j + c_4 j^2),$$

tandis que les solutions particulières linéairement indépendantes de (12) sont des fonctions de mailles $v_1(j) = 1$, $v_2(j) = 2^j$, $v_3(j) = j2^j$, $v_4(j) = j^2 2^j$.

4. Trouver la solution générale de l'équation d'ordre 4

$$y(j+4) + 8y(j+2) + 16y(j) = 0. \qquad (13)$$

L'équation caractéristique $q^4 + 8q^2 + 16 = (q^2+4)^2 = 0$ possède une racine complexe $q_1 = 2\left(\cos\frac{\pi}{2} + i\sin\frac{\pi}{2}\right)$ de multiplicité 2 et une racine qui lui est conjuguée $q_2 = 2\left(\cos\frac{\pi}{2} - i\sin\frac{\pi}{2}\right)$, également de multiplicité 2. Aussi la solution générale de l'équation (13) qui prend des valeurs réelles, a-t-elle la forme

$$y(j) = (c_1 + c_2 j)\, 2^j \cos\frac{\pi}{2} j + (c_3 + c_4 j)\, 2^j \sin\frac{\pi}{2} j.$$

Examinons encore deux exemples. Dans l'un on obtiendra la solution du problème de Cauchy pour une équation inhomogène de premier ordre, dans l'autre la solution du problème aux limites d'une équation homogène d'ordre 4.

5. Trouver la solution du problème suivant:

$$y(i+1) - ay(i) = f(i), \quad i \geqslant 0, \quad y(0) = y_0, \qquad (14)$$

où $a = \text{const}$. L'équation caractéristique $q - a = 0$ possède une racine unique $q_1 = a$. Aussi la solution générale de l'équation homogène prend-elle la forme: $\overline{\overline{y}}(i) = ca^i$, $c = \text{const}$. La solution particulière de l'équation inhomogène (14) sera recherchée en utilisant la méthode de variation de la constante. La formule (20) du § 2 fournit la solution particulière suivante de l'équation (14):

$$\overline{y}(i) = \sum_{k=0}^{i-1} a^{i-k-1} f(k) = \sum_{k=0}^{i-1} a^k f(i-k-1).$$

En vertu du théorème 3, la solution générale de l'équation inhomogène (14) a la forme

$$y(i) = ca^i + \sum_{k=0}^{i-1} a^k f(i-k-1).$$

En posant ici $i = 0$, on obtient (la somme disparaissant dans ce cas)

$y_0 = y(0) = c$. La solution du problème (14) est donc fournie par la formule

$$y(i) = y_0 a^i + \sum_{k=0}^{i-1} a^k f(i-k-1), \quad i \geqslant 0.$$

6. Cherchons maintenant la solution de l'équation d'ordre 4

$$y(j+2) - y(j+1) + 2y(j) - y(j-1) + y(j+2) = 0, \qquad 2 \leqslant j \leqslant N-2, \tag{15}$$

satisfaisant aux conditions aux limites suivantes:

$$\begin{gathered} 2y(2) - y(1) + y(0) = 2, \\ y(3) - y(2) + y(1) - y(0) = 0, \\ y(N-3) - y(N-2) + y(N-1) - y(N) = 0, \\ 2y(N-2) - y(N-1) + y(N) = 0. \end{gathered} \tag{16}$$

L'équation caractéristique

$$q^4 - q^3 + 2q^2 - q + 1 = (q^2 - q + 1)(q^2 + 1) = 0,$$

correspondant à (15), possède des racines complexes simples $q_1 = \cos\frac{\pi}{3} + i\sin\frac{\pi}{3}$, $q_2 = \cos\frac{\pi}{3} - i\sin\frac{\pi}{3}$, $q_3 = \cos\frac{\pi}{3} + i\sin\frac{\pi}{2}$, $q_4 = \cos\frac{\pi}{2} - i\sin\frac{\pi}{2}$; $i = \sqrt{-1}$. La solution générale de l'équation homogène (15), acquérant des valeurs réelles, a donc la forme

$$y(j) = c_1 \cos\frac{1}{3}\pi j + c_2 \sin\frac{1}{3}\pi j + c_3 \cos\frac{1}{2}\pi j + c_4 \sin\frac{1}{2}\pi j. \tag{17}$$

Dégageons maintenant de la solution générale (17) la solution vérifiant les conditions aux limites (16). A cette fin portons (17) dans (16) et l'on obtient le système suivant pour les constantes c_1, c_2, c_3 et c_4:

$$\begin{aligned} &\cos\frac{2\pi}{3}c_1 + \sin\frac{2\pi}{3}c_2 - c_3 - c_4 = 2, \\ &c_1 + 0\cdot c_2 + 0\cdot c_3 + 0\cdot c_4 = 0, \\ &\cos\frac{N\pi}{3}c_1 + \sin\frac{N\pi}{3}c_2 + 0\cdot c_3 + 0\cdot c_4 = 0, \\ &\cos\frac{(N-2)\pi}{3}c_1 + \sin\frac{(N-2)\pi}{3}c_2 - \left(\cos\frac{\pi N}{2} + \sin\frac{\pi N}{2}\right)c_3 + \\ &\qquad + \left(\cos\frac{\pi N}{2} - \sin\frac{\pi N}{2}\right)c_4 = 0. \end{aligned}$$

Le déterminant de ce système vaut $-2\sin\frac{N\pi}{3}\cos\frac{N\pi}{2}$ et est différent de zéro si N est pair mais n'est pas multiple de 3.

Dans ce cas, compte tenu de la parité de N, on obtient $c_1 = c_2 = 0$, $c_3 = c_4 = -1$. Donc si N est pair et n'est pas multiple de 3, la solution du problème aux limites (15), (16) existe et est fournie par la formule

$$y(j) = -\cos\frac{\pi j}{2} - \sin\frac{\pi j}{2}, \quad 0 \leqslant j \leqslant N.$$

Si N est impair ou est multiple de 3, la solution du problème (15), (16) est soit inexistante, soit non unique. Cet exemple constitue une illustration de la différence entre les problèmes aux limites dont la solution n'est pas toujours présente et le problème de Cauchy possédant une solution unique.

§ 4. Equations de second ordre à coefficients constants

1. Solution générale de l'équation homogène. Ce paragraphe est dévolu aux équations aux différences de second ordre à coefficients constants

$$a_2 y(j+2) + a_1 y(j+1) + a_0 y(j) = f(j), \quad a_0, a_2 \neq 0. \tag{1}$$

Cherchons d'abord la solution générale de l'équation homogène correspondante

$$a_2 y(j+2) + a_1 y(j+1) + a_0 y(j) = 0. \tag{2}$$

L'équation caractéristique $a_2 q^2 + a_1 q + a_0 = 0$ possède les racines

$$q_1 = \frac{-a_1 + \sqrt{a_1^2 - 4a_0 a_2}}{2a_2}, \quad q_2 = \frac{-a_1 - \sqrt{a_1^2 - 4a_0 a_2}}{2a_2}.$$

Selon la théorie générale des équations aux différences à coefficients constants, exposée au § 3, les solutions linéairement indépendantes de l'équation (2) sont les fonctions $v_1(j) = q_1^j$, $v_2(j) = q_2^j$ si $a_1^2 \neq 4a_0a_2$ et $v_1(j) = q_1^j$, $v_2(j) = jq_1^j$ si $a_1^2 = 4a_0a_2$. Dans la suite il sera commode d'utiliser d'autres solutions linéairement indépendantes

$$v_1(j) = \frac{q_2 q_1^j - q_1 q_2^j}{q_2 - q_1}, \quad v_2(j) = \frac{q_2^j - q_1^j}{q_2 - q_1}, \tag{3}$$

acquérant pour $j = 0$ et $j = 1$ les valeurs suivantes:

$$v_1(0) = 1, \quad v_1(1) = 0, \quad v_2(0) = 0, \quad v_2(1) = 1. \tag{4}$$

Il faut apparemment ne montrer que les fonctions (3) pour $a_1^2 = 4a_0a_2$ sont solutions de l'équation homogène. L'indépendance linéaire des fonctions (3) construites découle de la condition $\Delta_0(v_1, v_2) \neq 0$, où

$$\Delta_0(v_1, v_2) = \begin{vmatrix} v_1(0) & v_1(1) \\ v_2(0) & v_2(1) \end{vmatrix}.$$

En passant à la limite dans (3) avec q_2 tendant vers q_1, on obtient les fonctions $v_1(j) = -(j-1)\,q_1^j$. $v_2(j) = jq_1^{j-1}$ qui constituent en fait des solutions de l'équation homogène (2). Notons que les fonctions $v_1(j)$ et $v_2(j)$ de (3) prennent des valeurs réelles également dans le cas où les racines q_1 et q_2 sont complexes. Cela permet de ne pas étudier séparément le cas des racines complexes. Bref, la solution générale de l'équation homogène (2) peut être écrite sous la forme

$$\overline{\overline{y}}(j) = c_1 v_1(j) + c_2 v_2(j) = c_1 \frac{q_2 q_1^j - q_1 q_2^j}{q_2 - q_1} + c_2 \frac{q_2^j - q_1^j}{q_2 - q_1}, \tag{5}$$

où c_1 et c_2 sont des constantes arbitraires. Notons qu'en vertu de (4), il vient $\overline{\overline{y}}(0) = c_1$, $\overline{\overline{y}}(1) = c_2$.

Donnons un exemple. Il s'agit de trouver la solution générale de l'équation homogène

$$y(j+2) - 2xy(j+1) + y(j) = 0, \tag{6}$$

où x est un paramètre acquérant des valeurs réelles quelconques. Dans ce cas on a

$$q_1 = x + \sqrt{x^2-1}, \quad q_2 = \frac{1}{q_1}, \quad q_2 - q_1 = -2\sqrt{x^2-1}. \tag{7}$$

En portant (7) dans (5), on obtient la solution générale de l'équation (6) pour tout x sous la forme

$$y(j) = -\frac{(x+\sqrt{x^2-1})^{j-1} - (x+\sqrt{x^2-1})^{-(j-1)}}{2\sqrt{x^2-1}}\, y(0) + \\ + \frac{(x+\sqrt{x^2-1})^{j} - (x+\sqrt{x^2-1})^{-j}}{2\sqrt{x^2-1}}\, y(1). \tag{8}$$

En particulier, si $|x| \leqslant 1$, la formule (8) peut être écrite ainsi:

$$y(j) = -\frac{\sin(j-1)\arccos x}{\sin\arccos x}\, y(0) + \frac{\sin j\arccos x}{\sin\arccos x}\, y(1). \tag{9}$$

(Pour obtenir (9), on s'est servi de l'identité $x = \cos(\arccos x)$).

Profitons des résultats obtenus pour la résolution du problème posé au point 4, § 1 sur le calcul d'intégrales

$$I_k(\varphi) = \int_0^\pi \frac{\cos k\psi - \cos k\varphi}{\cos\psi - \cos\varphi}\, d\psi, \quad k = 0, 1, \ldots$$

On a montré alors que ce problème se réduit à la résolution du problème de Cauchy pour l'équation

$$I_{k+1} - 2\cos\varphi I_k + I_{k-1} = 0, \quad I_0 = 0, \quad I_1 = \pi. \tag{10}$$

Cette équation est un cas particulier de (6) avec $x = \cos\varphi$. Vu que $|x| \leqslant 1$, la solution générale de (10) est fournie sous la forme (9),

c'est-à-dire

$$I_k = -\frac{\sin (k-1)\varphi}{\sin \varphi} I_0 + \frac{\sin k\varphi}{\sin \varphi} I_1.$$

En y portant les données initiales de I_k, on obtient la solution du problème posé

$$I_k(\varphi) = \pi \frac{\sin k\varphi}{\sin \varphi}.$$

En qualité de second exemple examinons la solution du problème aux limites

$$\begin{gathered} y(j+1) - y(j) + y(j-1) = 0, \quad 1 \leqslant j \leqslant N-1, \\ y(0) = 1, \quad y(N) = 0. \end{gathered} \tag{11}$$

L'équation du problème (11) est également un cas particulier de (6) correspondant à la valeur $x = 1/2$. La formule (9) fournit la solution générale suivante de l'équation (11):

$$y(j) = \left(c_1 \sin \frac{(j-1)\pi}{3} + c_2 \sin \frac{j\pi}{3}\right) \Big/ \sin \frac{\pi}{3}.$$

Les constantes c_1 et c_2 s'obtiennent à partir des conditions aux limites de $y(j)$. Si N n'est pas multiple de 3, $c_1 = -1$, $c_2 = \sin \frac{1}{3}\pi (N - 1)/\sin \frac{1}{3} \pi N$, et la solution du problème (11) prend la forme

$$y(j) = \sin \frac{1}{3}(N-j)\pi/\sin \frac{1}{3} N\pi, \quad 0 \leqslant j \leqslant N.$$

Si N est multiple de 3, il n'y a pas de solution au problème (11).

2. Polynômes de Tchébychev. Revenons à l'équation (6). Examinons d'abord le problème de Cauchy suivant:

$$\begin{gathered} y(n+2) - 2xy(n+1) + y(n) = 0, \quad n \geqslant 0, \\ y(0) = 1, \quad y(1) = x. \end{gathered} \tag{12}$$

Notons que de (12) il découle

$$\begin{aligned} y(2) &= 2xy(1) - y(0) = 2x^2 - 1, \\ y(3) &= 2xy(2) - y(1) = 4x^3 - 3x, \end{aligned}$$

et, en général, $y(n)$ est un polynôme de degré n en x. Désignons ce polynôme par $T_n(x)$. En substituant $T_n(x)$ à $y(n)$ dans (12), on obtient la relation de récurrence à laquelle satisfait ce polynôme

$$\begin{gathered} T_{n+2}(x) = 2xT_{n+1}(x) - T_n(x), \quad n \geqslant 0, \\ T_0(x) = 1, \quad T_1(x) = x, \quad -\infty < x < \infty. \end{gathered} \tag{13}$$

D'autre part, la solution générale de l'équation (12) est fournie par la formule (8) pour tout x. Portons dans (8) les valeurs initiales

de $y(n)$, il vient alors

$$T_n(x) = \frac{(x+\sqrt{x^2-1})^n + (x+\sqrt{x^2-1})^{-n}}{2}. \tag{14}$$

En particulier, si $|x| \leqslant 1$, en posant $x = \cos(\arccos x)$, on obtient

$$T_n(x) = \cos(n \arccos x), \quad |x| \leqslant 1.$$

Bref, la solution du problème est trouvée. Cette solution est le polynôme $T_n(x)$, qui pour tout x se détermine par la formule (14) ou la formule

$$T_n(x) = \begin{cases} \cos(n \arccos x), & |x| \leqslant 1, \\ \frac{1}{2}[(x+\sqrt{x^2-1})^n + (x+\sqrt{x^2-1})^{-n}], & |x| \geqslant 1. \end{cases} \tag{15}$$

Le polynôme $T_n(x)$ est appelé *polynôme de Tchébychev de première espèce de degré n.*

Voyons maintenant un autre problème de Cauchy posé pour l'équation (6)

$$\begin{gathered} y(n+2) - 2xy(n+1) + y(n) = 0, \quad n \geqslant 0, \\ y(0) = 1, \quad y(1) = 2x. \end{gathered} \tag{16}$$

Ici aussi $y(n)$ est apparemment un polynôme du n-ième degré en x. Désignons-le par $U_n(x)$. Cherchons la forme explicite de $U_n(x)$. En portant les valeurs initiales de $y(n)$ dans (8), on obtient pour tout x:

$$\begin{aligned} U_n(x) &= \frac{2x(x+\sqrt{x^2-1})^n - (x+\sqrt{x^2-1})^{n-1}}{2\sqrt{x^2-1}} + \\ &\quad + \frac{(x+\sqrt{x^2-1})^{-(n-1)} - 2x(x+\sqrt{x^2-1})^{-n}}{2\sqrt{x^2-1}} = \\ &= \frac{(x+\sqrt{x^2-1})^{n+1} - (x+\sqrt{x^2-1})^{-(n+1)}}{2\sqrt{x^2-1}}. \end{aligned} \tag{17}$$

En particulier, si $|x| \leqslant 1$, alors

$$U_n(x) = \frac{\sin(n+1)\arccos x}{\sin \arccos x}.$$

Le polynôme $U_n(x)$ s'appelle *polynôme de Tchébychev de seconde espèce de degré n* et se définit par les formules

$$U_n(x) = \begin{cases} \dfrac{\sin(n+1)\arccos x}{\sin \arccos x}, & |x| \leqslant 1, \\ \dfrac{1}{2\sqrt{x^2-1}}[(x+\sqrt{x^2-1})^{n+1} - (x+\sqrt{x^2-1})^{-(n+1)}], & \\ \qquad |x| \geqslant 1. \end{cases} \tag{18}$$

A partir de (16) cherchons pour les polynômes $U_n(x)$ les relations de récurrence suivantes:

$$U_{n+2}(x) = 2xU_{n+1}(x) - U_n(x), \quad n \geqslant 0, \qquad U_0(x) = 1, \quad U_1(x) = 2x. \tag{19}$$

La formule (17) permet d'obtenir au lieu de (8) la représentation suivante de la solution générale de l'équation (6):

$$y(n) = -c_1U_{n-2}(x) + c_2U_{n-1}(x).$$

Cherchons encore une représentation de la solution générale de l'équation (6). Montrons que les fonctions $v_1(n) = T_n(x)$ et $v_2(n) = U_{n-1}(x)$ sont des solutions linéairement indépendantes de l'équation (6). En effet, il ne suffit que de montrer leur indépendance linéaire. Le déterminant

$$\Delta_0(v_1, v_2) = \begin{vmatrix} T_0(x) & T_1(x) \\ U_{-1}(x) & U_0(x) \end{vmatrix} = \begin{vmatrix} 1 & x \\ 0 & 1 \end{vmatrix} = 1$$

étant différent de zéro, l'assertion est vérifiée. La solution générale de l'équation (6) peut donc être représentée sous la forme

$$y(n) = c_1T_n(x) + c_2U_{n-1}(x), \tag{20}$$

où c_1 et c_2 sont des constantes arbitraires, tandis que les fonctions $T_n(x)$ et $U_n(x)$ sont définies par les formules (14) et (17) pour des x et n quelconques.

Pour conclure, donnons quelques relations aisément vérifiables, traduisant les liaisons entre les polynômes de Tchébychev $T_n(x)$ et $U_n(x)$, de même que les propriétés de ces polynômes. On a les formules suivantes:

$$T_n(x) = T_{-n}(x), \quad U_{-n}(x) = -U_{n-2}(x), \quad n \geqslant 0, \tag{21}$$

$$T_{in}(x) = T_i(T_n(x)), \quad U_{in-1}(x) = U_{i-1}(T_n(x)), \tag{22}$$

$$T_{2n}(x) = 2(T_n(x))^2 - 1, \tag{23}$$

$$T_{n-1}(x) - xT_n(x) = (1 - x^2)U_{n-1}(x), \tag{24}$$

$$U_{n-1}(x) - xU_n(x) = -T_{n+1}(x), \tag{25}$$

$$U_{n+i}(x) + U_{n-i}(x) = 2T_i(x)U_n(x). \tag{26}$$

A partir de (26), après substitution adéquate des indices i et n, il vient

$$U_{n+i-1}(x) + U_{n-i-1}(x) = 2T_i(x)U_{n-1}(x), \tag{27}$$

$$U_{n+i}(x) + U_{n-i-2}(x) = 2T_{i+1}(x)U_{n-1}(x). \tag{28}$$

En posant dans (26)-(28) $i = n$, on obtient

$$2T_n(x)U_n(x) = U_{2n}(x) + 1, \tag{29}$$

$$2T_n(x)U_{n-1}(x) = U_{2n-1}(x), \tag{30}$$

$$2T_{n+1}(x)U_{n-1}(x) = U_{2n}(x) - 1. \tag{31}$$

On a tenu compte ici des égalités (21) et de ce que $U_0(x) = 1$, $U_{-1}(x) = 0$. Si l'on pose dans (26) $n = 0$, il vient

$$2T_n(x) = U_n(x) - U_{n-2}(x). \tag{32}$$

3. Solution générale de l'équation inhomogène. Construisons maintenant la solution générale de l'équation inhomogène (1)

$$a_2 y(n+2) + a_1 y(n+1) + a_0 y(n) = f(n). \tag{33}$$

En vertu du théorème 3, la solution générale de l'équation (33) est la somme $y(n) = \bar{y}(n) + \bar{\bar{y}}(n)$, où $\bar{\bar{y}}(n)$ est la solution générale de l'équation homogène (2) et $\bar{y}(n)$ la solution particulière de l'équation inhomogène (33).

On a montré plus haut que les solutions linéairement indépendantes de l'équation (2) sont les fonctions

$$v_1(n) = \frac{q_2 q_1^n - q_1 q_2^n}{q_2 - q_1}, \quad v_2(n) = \frac{q_2^n - q_1^n}{q_2 - q_1}, \tag{34}$$

quant à la solution $\bar{\bar{y}}(n)$, elle est définie par la formule (5):

$$\bar{\bar{y}}(n) = c_1 v_1(n) + c_2 v_2(n).$$

Pour la recherche de la solution particulière $\bar{y}(n)$ de l'équation (33), servons-nous de la méthode de variation des constantes exposée au point 3, § 2. La formule (19) du § 2 fournit la solution $\bar{y}(n)$ sous la forme suivante:

$$\bar{y}(n) = \sum_{k=n_0}^{n-1} \left| \frac{\begin{matrix} v_1(k+1) & v_2(k+1) \\ v_1(k) & v_2(k) \end{matrix}}{\begin{matrix} v_1(k+1) & v_1(k+2) \\ v_2(k+1) & v_2(k+2) \end{matrix}} \right| \cdot \frac{f(k)}{a_2}.$$

Après des calculs peu laborieux, il vient

$$\bar{y}(n) = \sum_{k=n_0}^{n-2} \frac{q_2^{n-k-1} - q_1^{n-k-1}}{q_2 - q_1} \cdot \frac{f(k)}{a_2}, \quad n \neq n_0, \quad n_0 + 1$$

et

$$\bar{y}(n_0) = \bar{y}(n_0 + 1) = 0.$$

Par suite, la solution générale de l'équation inhomogène (33) prend la forme

$$y(n) = c_1 \frac{q_2 q_1^n - q_1 q_2^n}{q_2 - q_1} + c_2 \frac{q_2^n - q_1^n}{q_2 - q_1} + \sum_{k=n_0}^{n-2} \frac{q_2^{n-k-1} - q_1^{n-k-1}}{q_2 - q_1} \cdot \frac{f(k)}{a_2}, \tag{35}$$

où c_1 et c_2 sont des constantes arbitraires.

Au cas de résolution du problème de Cauchy, c'est-à-dire si l'on recherche la solution de l'équation (33) soumise aux conditions

$$y(n_0) = y_0, \quad y(n_0 + 1) = y_1, \tag{36}$$

on obtient alors à partir de (35) et (36) la représentation suivante de la solution de ce problème:

$$y(n) = y_0 \frac{q_2 q_1^{n-n_0} - q_1 q_2^{n-n_0}}{q_2 - q_1} + y_1 \frac{q^{n-n_0} \quad q_1^{n-n_0}}{q_2 - q_1} + \\ + \sum_{k=n_0}^{n-2} \frac{q_2^{n-k-1} - q_1^{n-k-1}}{q_2 - q_1} \cdot \frac{f(k)}{a_2}. \tag{37}$$

Cherchons maintenant la solution du premier problème aux limites d'une équation aux différences d'ordre deux à coefficients constants. Il est commode d'écrire ce problème sous la forme suivante:

$$\begin{gathered} a_2 y(n+1) + a_1 y(n) + a_0 y(n-1) = -f(n), \quad 1 \leqslant n \leqslant N-1, \\ y(0) = \mu_1, \quad y(N) = \mu_2. \end{gathered} \tag{38}$$

Cette écriture diffère de celle de (33) par le déplacement de l'indice n, aussi, en utilisant (35), obtient-on la formule suivante pour la solution générale de l'équation (38):

$$y(n) = c_1 \frac{q_2 q_1^n - q_1 q_2^n}{q_2 - q_1} + c_2 \frac{q_2^n - q_1^n}{q_2 - q_1} - \sum_{k=1}^{n-1} \frac{q_2^{n-k} - q_1^{n-k}}{q_2 - q_1} \cdot \frac{f(k)}{a_2}. \tag{39}$$

Déterminons les constantes c_1 et c_2 à partir de la condition obligeant la solution (39) de prendre les valeurs données $y(0) = \mu_1$ et $y(N) = \mu_2$ pour $n = 0$ et $n = N$ respectivement. En omettant les calculs peu compliqués, on obtient la formule suivante de la solution du problème aux limites (38):

$$y(n) = \frac{(q_1 q_2)^n (q_2^{N-n} - q_1^{N-n})}{q_2^N - q_1^N} \mu_1 + \frac{q_2^n - q_1^n}{q_2^N - q_1^N} \mu_2 + \\ + \sum_{k=1}^{n-1} \frac{(q_1 q_2)^{n-k} (q_2^{N-n} - q_1^{N-n})(q_2^k - q_1^k)}{(q_2 - q_1)(q_2^N - q_1^N)} \cdot \frac{f(k)}{a_2} + \\ + \sum_{k=n}^{N-1} \frac{(q_2^{N-k} - q_1^{N-k})(q_2^n - q_1^n)}{(q_2 - q_1)(q_2^N - q_1^N)} \cdot \frac{f(k)}{a_2}. \tag{40}$$

Notons que la solution du problème aux limites (38) n'existe pas qu'au cas où $q_1^N = q_2^N$, mais $q_1 \neq q_2$.

Examinons maintenant les cas particuliers d'application de la formule (40). Supposons qu'il s'agit de résoudre le premier problème

aux limites pour l'équation

$$y(n+1)-2xy(n)+y(n-1)=-f(n),\quad 1\leqslant n\leqslant N-1,$$
$$y(0)=\mu_1,\quad y(N)=\mu_2. \tag{41}$$

On a trouvé plus haut les racines q_1 et q_2 de l'équation caractéristique correspondant à (41)

$$q_1=x+\sqrt{x^2-1},\quad q_2=x-\sqrt{x^2-1}=1/q_1.$$

En portant ces valeurs dans (40) et compte tenu de la formule (17) établie pour le polynôme $U_n(x)$, on obtient la solution du problème (41) sous la forme

$$y(n)=\frac{U_{N-n-1}(x)}{U_{N-1}(x)}\left[\mu_1+\sum_{k=1}^{n-1}U_{k-1}(x)f(k)\right]+$$
$$+\frac{U_{n-1}(x)}{U_{N-1}(x)}\left[\mu_2+\sum_{k=n}^{N-1}U_{N-k-1}(x)f(k)\right]. \tag{42}$$

La solution existe et est fournie par la formule (42), si la condition $x\neq\cos\frac{k\pi}{N}$, $k=1,2,\ldots,N-1$ est remplie.

Revenons à l'équation (38). Si $a_0a_2>0$, la solution (40) de ce problème peut être écrite sous une forme plus condensée que (40). En effet, écrivons les racines

$$q_1=\frac{1}{2a_2}\left[-a_1+\sqrt{a_1^2-4a_0a_2}\right],\quad q_2=\frac{1}{2a_2}\left[-a_1-\sqrt{a_1^2-4a_0a_2}\right]$$

de l'équation caractéristique correspondant à (38) sous la forme suivante

$$q_1=\rho\left(x+\sqrt{x^2-1}\right),\quad q_2=\rho\left(x-\sqrt{x^2-1}\right), \tag{43}$$

où

$$\rho=\sqrt{\frac{a_0}{a_2}},\quad x=-\frac{a_1}{2\sqrt{a_0a_2}}. \tag{44}$$

Portons (43) dans (40), compte tenu de la formule (17). On obtient la solution du problème (38) pour le cas où $a_0a_2>0$ sous la forme

$$y(n)=\frac{U_{N-n-1}(x)}{U_{N-1}(x)}\rho^n\left[\mu_1+\sum_{k=1}^{n-1}\frac{U_{k-1}(x)}{\rho^{k-1}}\cdot\frac{f(k)}{a_0}\right]+$$
$$+\frac{U_{n-1}(x)}{U_{N-1}(x)}\cdot\frac{1}{\rho^{N-n}}\left[\mu_2+\sum_{k=n}^{N-1}\rho^{N-k-1}U_{N-k-1}(x)\frac{f(k)}{a_0}\right],$$

où ρ et x sont définis dans (44). La solution du problème (38) pour le cas de $a_0a_2 > 0$ existe, si la condition $a_1 + 2\sqrt{a_0a_2}\cos\frac{k\pi}{N} \neq 0$ est remplie, $k = 1, 2, \ldots, N-1$.

Etudions maintenant le premier problème aux limites pour une équation vectorielle triponctuelle à coefficients constants

$$\begin{aligned} &Y_{n-1} - CY_n + Y_{n+1} = -F_n, \quad 1 \leqslant n \leqslant N-1, \\ &Y_0 = F_0, \quad Y_N = F_N, \end{aligned} \tag{45}$$

où Y_n et F_n sont des vecteurs et C une matrice carrée. Il est aisé de s'assurer que la solution générale de l'équation inhomogène (45) prend la forme

$$Y_n = U_{n-2}\left(\frac{1}{2}C\right)C_1 + U_{n-1}\left(\frac{1}{2}C\right)C_2 - \sum_{k=1}^{n-1} U_{n-k-1}\left(\frac{1}{2}C\right)F_k,$$

où C_1 et C_2 sont des vecteurs arbitraires et $U_n(X)$ est le polynôme matriciel de la matrice X, défini suivant les formules de récurrence (19).

Si la matrice C est telle que $U_{N-1}\left(\frac{1}{2}\right)C$ devient une matrice non dégénérée, la solution du problème aux limites (45) se détermine alors par une formule analogue à la formule (42)

$$\begin{aligned} Y_n = U_{N-1}^{-1}\left(\frac{1}{2}C\right)U_{N-n-1}\left(\frac{1}{2}C\right)\left[F_0 + \sum_{k=1}^{n-1} U_{k-1}\left(\frac{1}{2}C\right)F_k\right] + \\ + U_{N-1}^{-1}\left(\frac{1}{2}C\right)U_{n-1}\left(\frac{1}{2}C\right)\left[F_N + \sum_{k=n}^{N-1} U_{N-k-1}\left(\frac{1}{2}C\right)F_k\right]. \end{aligned} \tag{46}$$

On montrera plus loin qu'au problème (45) se réduit le problème de Dirichlet au sens de différences finies pour l'équation de Poisson dans un rectangle.

Remarquons, en guise de conclusion, que la condition de l'existence de la solution du problème (45) peut être formulée de la façon suivante : la solution existe et se détermine à l'aide de la formule (46) si les nombres $\cos\frac{k\pi}{N}$, $k = 1, 2, \ldots, N-1$, ne constituent pas des valeurs propres de la matrice C.

§ 5. Problèmes de différences de valeurs propres

1. Premier problème aux limites de valeurs propres. Dans le chapitre IV on abordera l'étude de la méthode de la séparation des variables, qui est utilisée à des fins de recherche des solutions aux problèmes aux limites discrets pour des équations elliptiques dans

un rectangle. Sous ce rapport s'élève la nécessité de représenter les fonctions de mailles cherchées sous forme d'un développement en fonctions propres du problème discret correspondant. Dans ce paragraphe on étudiera les problèmes de différences sur les valeurs propres pour le plus simple des opérateurs de différences de second ordre donné sur un maillage régulier.

Formulons le premier problème aux limites. Soit sur le segment $[0, l]$ un maillage régulier $\bar{\omega} = \{x_i = ih, i = 0, 1, \ldots, N, hN = l\}$ avec pas h. Il s'agit de trouver les valeurs du paramètre λ (valeurs propres) pour lesquelles on a une solution non triviale $\mu(x_i)$ (fonctions propres) du problème de différences suivant:

$$y_{\bar{x}x} + \lambda y = 0, \quad x \in \omega, \quad y(0) = y(l) = 0, \tag{1}$$

où

$$y_{\bar{x}x,\, i} = \frac{y(i+1) - 2y(i) + y(i-1)}{h^2}, \quad y(i) = y(x_i).$$

Cherchons la solution du problème (1). Pour cela écrivons (1) sous forme de problème aux limites pour une équation aux différences d'ordre deux

$$y(i+1) - 2\left(1 - \frac{h^2\lambda}{2}\right) y(i) + y(i-1) = 0, \quad 1 \leqslant i \leqslant N-1,$$
$$y(0) = y(N) = 0. \tag{2}$$

Au point 1 du § 4 on a montré que la solution générale de l'équation (2) prend la forme (voir formule (20) du § 4) $y(i) = c_1 T_i(z) + c_2 U_{i-1}(z)$, où c_1 et c_2 sont des constantes arbitraires, tandis que z désigne ici l'expression

$$z = 1 - h^2\lambda/2. \tag{3}$$

Les constantes c_1 et c_2 se déterminent à partir des conditions aux limites

$$y(0) = c_1 = 0, \quad y(N) = c_2 U_{N-1}(z) = 0. \tag{4}$$

Ici et dans la suite on utilise les formules (15) et (18) du § 4, où sont définis les polynômes de Tchébychev des première et seconde espèces, de même que les formules (21)-(32) du même paragraphe.

Comme on se propose de rechercher la solution non triviale de (1), on a $c_2 \neq 0$, et à partir de (4) on obtient la condition $U_{N-1}(z) = 0$ qui, une fois satisfaite, nous donne la solution du problème (1) sous la forme $y_i = c_2 U_{i-1}(z)$.

Vu que les nombres $z_k = \cos\frac{k\pi}{N}$, $k = 1, 2, \ldots, N-1$, sont des racines du polynôme $U_{N-1}(z)$, à partir de (3) on déduit les valeurs propres du problème (1)

$$\lambda_k = \frac{4}{h^2}\sin^2\frac{k\pi}{2N} = \frac{4}{h^2}\sin^2\frac{k\pi h}{2l}, \quad k = 1, 2, \ldots, N-1. \tag{5}$$

A chaque valeur propre λ_k correspond la solution non nulle du problème (1)

$$y_k(i) = c_2 U_{i-1}(z_k) = \bar{c}_k \sin\frac{k\pi i}{N} = \bar{c}_k \sin\frac{k\pi x_i}{l},$$
$$0 \leqslant i \leqslant N \quad \left(c_2 = \bar{c}_k \sin\frac{k\pi}{N}\right). \qquad (6)$$

Déterminons le produit scalaire des fonctions de mailles associées à $\bar{\omega}$ de la façon suivante:

$$(u, v) = \sum_{i=1}^{N-1} u(i)\, v(i)\, h + 0{,}5h\,[u(0)\, v(0) + u(N)\, v(N)].$$

Déterminons maintenant la constante $\bar{c}_k$ de (6) de manière qu les fonctions $y_k(i)$ aient la norme égale à l'unité, c'est-à-dir $(y_k, y_k) = 1$.

Des calculs élémentaires donnent $\bar{c}_k = \sqrt{2/l}$. En portant la valeur trouvée de $\bar{c}_k$ dans (6), on obtient les fonctions propres $\mu_k(i)$ du problème (1)

$$\mu_k(i) = \sqrt{\frac{2}{l}} \sin\frac{k\pi i}{N} = \sqrt{\frac{2}{l}} \sin\frac{k\pi x_i}{l}, \qquad (7)$$
$$i = 0, 1, \ldots, N, \quad k = 1, 2, \ldots, N-1.$$

Bref, le problème (1) est résolu et la solution est fournie par (5) et (7).

Enumérons les principales propriétés des fonctions propres et des valeurs propres du premier problème aux limites (1).

1) Les fonctions propres sont orthonormées:

$$(\mu_k, \mu_m) = \delta_{km}, \quad \delta_{km} = \begin{cases} 1, & k = m, \\ 0, & k \neq m. \end{cases}$$

2) Pour toute fonction de maille $f(i)$ donnée sur les nœuds internes du maillage $\bar{\omega}$, c'est-à-dire pour $1 \leqslant i \leqslant N-1$, on a le développement

$$f(i) = \frac{2}{N} \sum_{k=1}^{N-1} \varphi_k \sin\frac{k\pi i}{N}, \quad i = 1, 2, \ldots, N-1, \qquad (8)$$

où

$$\varphi_k = \sum_{i=1}^{N-1} f(i) \sin\frac{k\pi i}{N}, \quad k = 1, 2, \ldots, N-1. \qquad (9)$$

Eclairons cette assertion. Soit $f(i)$ une fonction de maille quelconque donnée sur ω (ou sur $\bar{\omega}$ et devenant nulle pour $i = 0$ et $i = N$).

Développons-la en fonctions propres

$$f(i)=\sum_{i=1}^{N-1} f_k\mu_k(i)=\sum_{k=1}^{N-1}\sqrt{\frac{2}{l}}\, f_k \sin\frac{k\pi i}{N}, \tag{10}$$

où f_k est le coefficient de Fourier de la fonction $f(i)$. En multipliant scalairement (10) par $\mu_m(i)$ et profitant de l'orthonormalité des fonctions propres, on obtient les coefficients de Fourier

$$f_m=\sum_{k=1}^{N-1} f_k(\mu_k,\ \mu_m)=(f,\ \mu_m)=\sum_{i=1}^{N-1}\sqrt{\frac{2}{l}}\, f(i)\sin\frac{\pi k i}{N}\, h.$$

La liaison des formules obtenues avec (8)-(9) s'établit aisément en remarquant que $f_m=\frac{\sqrt{2l}}{N}\varphi_m$.

Le développement de (8), (9) est commode par le fait que pour le calcul de l'image Fourier de la fonction $f(i)$, ainsi que pour le rétablissement de la fonction primitive d'après son image, il faut calculer une somme du même type. L'algorithme du calcul rapide des sommes de cette sorte sera étudié au ch. IV.

3) Pour les valeurs propres sont vérifiées les inégalités suivantes

$$\frac{8}{l^2}\leqslant\frac{4}{h^2}\sin^2\frac{\pi}{2N}=\lambda_1\leqslant\lambda_k\leqslant\lambda_{N-1}=\frac{4}{h^2}\cos^2\frac{\pi}{2N},\quad 1\leqslant k\leqslant N-1.$$

2. Second problème aux limites. Etudions maintenant le second problème aux limites de valeurs propres

$$\begin{gathered} y_{\bar{x}x}+\lambda y=0,\quad x\in\omega,\\ \frac{2}{h}y_x+\lambda y=0,\quad x=0,\quad -\frac{2}{h}y_{\bar{x}}+\lambda y=0,\quad x=l. \end{gathered} \tag{11}$$

Cherchons la solution du problème (11). En répartissant les différences dans (11) suivant les points, on aboutit au problème

$$\begin{gathered} y(i+1)-2zy(i)+y(i-1)=0,\quad 1\leqslant i\leqslant N-1,\\ y(1)-zy(0)=0,\quad y(N-1)-zy(N)=0, \end{gathered} \tag{12}$$

où $z=1-\lambda h^2/2$. De la solution générale de l'équation (12) $y(i)=c_1T_i(z)+c_2U_{i-1}(z)$ séparons la solution satisfaisant aux conditions aux limites posées. En utilisant la formule (24) du § 4, on obtient

$$y(1)-zy(0)=c_1z+c_2-c_1z=c_2=0,\quad c_2=0,$$

de même que

$$y(N-1)-zy(N)=c_1(T_{N-1}(z)-zT_N(z))=$$
$$=c_1(1-z^2)U_{N-1}(z)=0.$$

Puisque $c_1 \neq 0$, il s'ensuit de ce qui précède que

$$z_k = \cos\frac{k\pi}{N}, \quad k = 0, 1, \ldots, N,$$

et, par suite, les valeurs propres du problème (12) sont:

$$\lambda_k = \frac{4}{h^2}\sin^2\frac{k\pi}{2N} = \frac{4}{h^2}\sin^2\frac{k\pi h}{2l}, \quad k = 0, 1, \ldots, N. \qquad (13)$$

De plus, à chaque λ_k correspond une solution non nulle du problème (11)

$$y_k(i) = c_k T_i(z_k) = c_k \cos\frac{k\pi i}{N}, \quad 0 \leqslant i \leqslant N.$$

Tirons les constantes c_k de la condition $(y_k, y_k) = 1$, dont le produit scalaire est défini plus haut. Des calculs directs montrent que

$$c_k = \sqrt{2/l}, \quad k = 1, 2, \ldots, N-1, \quad c_k = \sqrt{1/l}, \quad k = 0, N.$$

Donc les fonctions propres normées du problème (11) sont les fonctions

$$\begin{aligned} \mu_k(i) &= \sqrt{\frac{2}{l}}\cos\frac{k\pi i}{N} = \sqrt{\frac{2}{l}}\cos\frac{k\pi x_i}{l}, \quad 1 \leqslant k \leqslant N-1, \\ \mu_k(i) &= \sqrt{\frac{1}{l}}\cos\frac{k\pi i}{N} = \sqrt{\frac{1}{l}}\cos\frac{k\pi x_i}{l}, \quad k = 0, N, \end{aligned} \qquad (14)$$

données sur le maillage $\bar{\omega}$. Notons que la fonction propre correspondant à la solution propre nulle $\lambda_0 = 0$ est la constante $\mu_0(i) = \sqrt{1/l}$.

Formulons les propriétés des fonctions propres et des valeurs propres du second problème aux limites (11).

1) Les fonctions propres sont orthonormées: $(\mu_k, \mu_m) = \delta_{km}$.

2) Pour toute fonction de maille $f(i)$ donnée sur $\bar{\omega}$ on a le développement

$$f(i) = \frac{2}{N}\sum_{k=0}^{N} \rho_k \varphi_k \cos\frac{k\pi i}{N}, \quad i = 0, 1, \ldots, N, \qquad (15)$$

où

$$\varphi_k = \sum_{i=0}^{N} \rho_i f(i) \cos\frac{k\pi i}{N}, \quad k = 0, 1, \ldots, N, \qquad (16)$$

$$\rho_i = \begin{cases} 1, & 1 \leqslant i \leqslant N-1, \\ 0{,}5, & i = 0, N. \end{cases} \qquad (17)$$

Les formules (15) et (16) sont des modifications du développement traditionnel de $f(i)$ en fonctions propres $\mu_k(i)$

$$f(i)=\sum_{k=0}^{N} f_k\mu_k(i), \quad f_k=(f, \mu_k)$$

au moyen des substitutions suivantes:

$$f_k=\begin{cases} \dfrac{\sqrt{2l}}{N}\varphi_k, & 1\leqslant k\leqslant N-1, \\ \dfrac{1}{N}\sqrt{l}\,\varphi_k, & k=0,\ N. \end{cases}$$

3) Pour les valeurs propres se vérifient les inégalités

$$0=\lambda_0\leqslant\lambda_k\leqslant\lambda_N, \quad 0\leqslant k\leqslant N.$$

3. Problème aux limites mixte. Voyons maintenant le problème de valeurs propres quand à un bout du segment $[0, l]$ est imposée la condition aux limites de première espèce et à l'autre, de seconde espèce, par exemple:

$$\begin{gathered} y_{\bar{x}x}+\lambda y=0, \quad x\in\omega, \\ y(0)=0, \quad -\frac{2}{h}y_{\bar{x}}+\lambda y=0, \quad x=l. \end{gathered} \tag{18}$$

Un tel problème sera appelé *problème aux limites mixte*.

Cherchons la solution du problème (18). Le problème de l'équation aux différences d'ordre deux correspondant à (18) a la forme

$$\begin{aligned} &y(i+1)-2zy(i)+y(i-1)=0, \quad 1\leqslant i\leqslant N-1, \\ &y(0)=0, \quad y(N-1)-zy(N)=0, \end{aligned}$$

où $z=1-0{,}5\lambda h^2$. Séparons de la solution générale de cette équation

$$y(i)=c_1T_i(z)+c_2U_{i-1}(z)$$

la solution satisfaisant les conditions aux limites données. En utilisant (25) du § 4, il vient

$$y(0)=c_1=0,$$
$$y(N-1)-zy(N)=c_2(U_{N-2}(z)-zU_{N-1}(z))=-c_2T_N(z)=0.$$

Vu que $c_2\neq 0$, on obtient de cette expression $T_N(z_k)=0$, où $z_k=\cos\dfrac{(2k-1)\pi}{2N}$, $k=1, 2, \ldots, N$ et, par suite, les valeurs propres du problème (18) sont les nombres

$$\lambda_k=\frac{4}{h^2}\sin^2\frac{(2k-1)\pi}{4N}=\frac{4}{h^2}\sin^2\frac{(2k-1)\pi h}{4l}, \quad k=1, 2, \ldots, N. \tag{19}$$

Les fonctions propres normées du problème (18) qui correspondent aux valeurs propres λ_k sont

$$\mu_k(i) = \sqrt{\frac{2}{l}} \sin \frac{(2k-1)\pi i}{2N} = \sqrt{\frac{2}{l}} \sin \frac{(2k-1)\pi x_i}{2l}, \quad k = 1, 2, \ldots, N. \quad (20)$$

Formulons les propriétés des fonctions propres et des valeurs propres du problème aux limites mixte (18).

1) Les fonctions propres sont orthonormées: $(\mu_k, \mu_m) = \delta_{km}$.

2) Pour toute fonction de maille $f(i)$ donnée sur $\omega^+ = \{x_i = ih, 1 \leqslant i \leqslant N\}$ (ou sur $\overline{\omega}$ et devenant nulle pour $i = 0$) se vérifie le développement

$$f(i) = \frac{2}{N} \sum_{k=1}^{N} \varphi_k \sin \frac{(2k-1)\pi i}{2N}, \quad i = 1, 2, \ldots, N, \quad (21)$$

où

$$\varphi_k = \sum_{i=1}^{N} \rho_i f(i) \sin \frac{(2k-1)\pi i}{2N}, \quad k = 1, 2, \ldots, N, \quad (22)$$

(ρ_i est déterminé dans (17)).

3) Pour les valeurs propres se vérifient les inégalités

$$\frac{8}{(2+\sqrt{2})\,l^2} \leqslant \frac{4}{h^2} \sin^2 \frac{\pi}{2N} = \lambda_1 \leqslant \lambda_k \leqslant \lambda_N = \frac{4}{h^2} \cos^2 \frac{\pi}{4N}, \quad 1 \leqslant k \leqslant N.$$

Si pour l'équation (18) la condition aux limites de première espèce est imposée au bout droit du segment $[0, l]$, c'est-à-dire qu'est donné le problème

$$\begin{gathered} y_{\bar{x}x} + \lambda y = 0, \quad x \in \omega, \\ \frac{2}{h} y_x + \lambda y = 0, \quad x = 0; \quad y(l) = 0, \end{gathered} \quad (23)$$

les valeurs propres se déterminent alors par la formule (19), tandis que les fonctions propres normées sont

$$\mu_k(i) = \sqrt{\frac{2}{l}} \sin \frac{(2k-1)(N-i)\pi}{2N} = \sqrt{\frac{2}{l}} \sin \frac{(2k-1)\pi(l-x_i)}{2l}, \quad k = 1, 2, \ldots, N.$$

On peut formuler la proposition suivante. *Pour toute fonction de maille $f(i)$ donnée sur $\omega^- = \{x_i = ih, i = 0, 1, \ldots, N-1, hN =$*

$= l\}$ *(ou sur* $\bar{\omega}$ *et devenant nulle pour* $i = N$*) se vérifie le développement*

$$f(N-i) = \frac{2}{N} \sum_{k=1}^{N} \varphi_k \sin \frac{(2k-1)\pi i}{2N}, \quad i = 1, 2, \ldots, N, \qquad (24)$$

où

$$\varphi_i = \sum_{i=1}^{N} \rho_{N-i} f(N-i) \sin \frac{(2k-1)\pi i}{2N}, \quad k = 1, 2, \ldots, N, \qquad (25)$$

(ρ_i *est déterminé dans* (17)).

Remarquons que les fonctions propres construites pour le problème (23) sont également orthonormées:

$$(\mu_k, \mu_m) = \delta_{km}.$$

4. **Problème aux limites périodique.** Posons que sur un maillage $\Omega = \{x_i = ih, i = 0, \pm 1, \pm 2, \ldots\}$, introduit sur une droite $-\infty < x < \infty$, on recherche une solution périodique non triviale de période N du problème suivant de valeurs propres:

$$\begin{aligned} &y_{\bar{x}x} + \lambda y = 0, \quad x \in \Omega, \\ &y(i+N) = y(i), \quad i = 0, \pm 1, \pm 2, \ldots, \quad h = l/N. \end{aligned} \qquad (26)$$

Vu que la solution est périodique, il suffit de la trouver pour $i = 0, 1, \ldots, N-1$. En répartissant (26) suivant les points $i = 0, 1, \ldots, N-1$ et tenant compte de ce que $y(-1) = y(N-1)$, $y(0) = y(N)$, on obtient le problème suivant:

$$\begin{aligned} &y(i+1) - 2zy(i) + y(i-1) = 0, \quad 0 \leqslant i \leqslant N-1, \qquad (27) \\ &y(0) = y(N), \quad y(-1) = y(N-1), \end{aligned}$$

où $z = 1 - 0{,}5\lambda h^2$.

Cherchons la solution du problème (27). Imposons à la solution générale

$$y(i) = c_1 T_i(z) + c_2 U_{i-1}(z)$$

les conditions aux limites. Compte tenu des propriétés des polynômes de Tchébychev, on obtient le système suivant permettant de déterminer les constantes c_1 et c_2:

$$\begin{aligned} &c_1(1 - T_N(z)) - c_2 U_{N-1}(z) = 0, \qquad (28) \\ &c_1(T_{N-1}(z) - z) + c_2(1 + U_{N-2}(z)) = 0. \end{aligned}$$

Ce système a une solution non nulle seulement et rien que si son déterminant est nul. Calculons-le en recourant, à des fins de trans-

formation, aux formules (25), (29) et (31) du § 4. On obtient

$$(1 - T_N(z))(1 + U_{N-2}(z)) + (T_{N-1}(z) - z)\,U_{N-1}(z) =$$
$$= 1 + U_{N-2}(z) - zU_{N-1}(z) - T_N(z) + T_{N-1}(z)\,U_{N-1}(z) -$$
$$- T_N(z)\,U_{N-2}(z) = 2\,[1 - T_N(z)] = 0.$$

Il en découle que pour $z = z_k$, où

$$z_k = \cos\frac{2k\pi}{N}, \quad k = 0,\ 1,\ \ldots,\ N-1, \tag{29}$$

le système (28) possède une solution non nulle. Donc les valeurs propres du problème (26) sont

$$\lambda_k = \frac{4}{h^2}\sin^2\frac{k\pi}{N} = \frac{4}{h^2}\sin^2\frac{k\pi h}{l}, \quad k = 0,\ 1,\ \ldots,\ N-1. \tag{30}$$

Cherchons maintenant la solution du système (28). Puisqu'on a les égalités

$$T_{N-1}(z_k) = z_k, \quad 0 \leqslant k \leqslant N-1,$$
$$U_{N-2}(z_k) = \begin{cases} N-1, & k = 0,\ N/2, \\ -1, & k \neq 0,\ N/2, \end{cases}$$
$$U_{N-1}(z_k) = \begin{cases} N, & k = 0, \\ -N, & k = N/2, \\ 0, & k \neq 0,\ N/2, \end{cases}$$

en portant (29) dans (28), on obtient la solution suivante du système (28):

a) pour $k = 0$ et $k = N/2$, on a $c_2 = 0$, $c_1 = c_1^{(k)} \neq 0$;

b) pour $k \neq 0$, $k \neq N/2$, $0 < k \leqslant N-1$, les constantes $c_1 = c_1^{(k)}$, $c_2 = c_2^{(k)}$ sont quelconques, mais ne sont pas nulles simultanément. De là on obtient que les fonctions

$$\begin{aligned} y_k(i) &= c_1^{(k)}\cos\frac{2k\pi i}{N}, \quad k = 0,\ N/2, \\ y_k(i) &= c_1^{(k)}\cos\frac{2k\pi i}{N} + c_2^{(k)}\sin\frac{2k\pi i}{N}, \quad 1 \leqslant k \leqslant N-1, \\ &\qquad k \neq 0,\ \frac{N}{2} \end{aligned} \tag{31}$$

sont des solutions du problème (27) correspondant à la valeur propre λ_k. Notons qu'au cas où $k \neq 0$, $N/2$ les formules (31) déterminent en fait deux fonctions linéairement indépendantes $c_1^{(k)}\cos\frac{2k\pi i}{N}$ et $c_2^{(k)}\sin\frac{2k\pi i}{N}$, dont chacune constitue une solution du problème (27) et correspond à la valeur propre λ_k.

Construisons maintenant les fonctions propres normées du problème (26). Notons que pour les fonctions de mailles périodiques le produit scalaire introduit précédemment peut être écrit de la façon suivante:

$$(u, v)_{\bar{\omega}} = \sum_{i=1}^{N-1} u(i)\, v(i)\, h + 0{,}5h\, [u(0)\, v(0) + u(N)\, v(N)] = \\ = \sum_{i=0}^{N-1} u(i)\, v(i)\, h.$$

Examinons deux cas. Soit d'abord N pair. De (31) on obtient que les fonctions propres correspondant à λ_0 et $\lambda_{N/2}$ sont

$$\mu_k(i) = \sqrt{\frac{1}{l}} \cos \frac{2k\pi i}{N}, \quad k = 0,\ \frac{N}{2}. \tag{32}$$

Ensuite remarquons qu'à partir de (30) s'ensuit l'égalité

$$\lambda_{N-k} = \frac{4}{h^2} \sin^2 \frac{(N-k)\pi}{N} = \frac{4}{h^2} \sin^2 \frac{k\pi}{N} = \lambda_k, \\ k = 1,\ 2,\ \ldots,\ \frac{N}{2} - 1.$$

En choisissant en qualité de fonctions propres, correspondant à la valeur propre λ_k, la fonction

$$\mu_k(i) = \sqrt{\frac{2}{l}} \cos \frac{2k\pi i}{N}, \quad 1 \leqslant k \leqslant \frac{N}{2} - 1$$

et la fonction

$$\mu_{N-k}(i) = \sqrt{\frac{2}{l}} \sin \frac{2k\pi i}{N}, \quad 1 \leqslant k \leqslant \frac{N}{2} - 1,$$

correspondant à la valeur $\lambda_{N-k} = \lambda_k$, on obtient avec (32) le système complet des fonctions propres du problème (26). Bref, les valeurs propres de λ_k sont celles définies dans (30), tandis que les fonctions propres du problème (26) sont données par les formules

$$\begin{aligned} \mu_k(i) &= \sqrt{\frac{1}{l}} \cos \frac{2k\pi i}{N}, & k &= 0,\ \frac{N}{2}, \\ \mu_k(i) &= \sqrt{\frac{2}{l}} \cos \frac{2k\pi i}{N}, & 1 &\leqslant k \leqslant \frac{N}{2} - 1, \\ \mu_k(i) &= \sqrt{\frac{2}{l}} \sin \frac{2(N-k)\pi i}{N}, & \frac{N}{2} + 1 &\leqslant k \leqslant N - 1 \end{aligned} \tag{33}$$

au cas de N pair.

Notons les principales propriétés des fonctions propres et des valeurs propres du problème aux limites périodique (26).

1) Les fonctions propres sont orthonormées.

2) Toute fonction de maille $f(i)$ périodique de période N donnée sur le maillage Ω peut être représentée sous la forme

$$f(i)=\frac{2}{N}\sum_{k=0}^{N/2}\rho_k\varphi_k\cos\frac{2k\pi i}{N}+\frac{2}{N}\sum_{k=N/2+1}^{N-1}\varphi_k\sin\frac{2(N-k)\pi i}{N},\qquad(34)$$

où

$$\begin{aligned}\varphi_k&=\sum_{i=0}^{N-1}\rho_k f(i)\cos\frac{2k\pi i}{N}, && 0\leqslant k\leqslant\frac{N}{2},\\ \varphi_k&=\sum_{i=0}^{N-1}f(i)\sin\frac{2(N-k)\pi i}{N}, && \frac{N}{2}+1\leqslant k\leqslant N-1,\end{aligned}\qquad(35)$$

$$\rho_k=\begin{cases}1, & k\neq 0,\ N/2,\\ 1/\sqrt{2}, & k=0,\ N/2.\end{cases}\qquad(36)$$

Les formules (34)-(36) se déduisent du développement de la fonction $f(i)$ en fonctions propres $\mu_k(i)$:

$$f(i)=\sum_{k=0}^{N-1}f_k\mu_k(i),\qquad f_k=(f,\ \mu_k)$$

avec la substitution $f_k=\frac{\sqrt{2l}}{N}\varphi_k$.

3) Pour les valeurs propres se vérifient les inégalités

$$0=\lambda_0\leqslant\lambda_k\leqslant\lambda_{N/2}=\frac{4}{h^2},\qquad 0\leqslant k\leqslant N-1.$$

Voyons maintenant le cas où N est impair. Dans ce cas les valeurs propres du problème (26) se déterminent par les formules (30), de plus on a $\lambda_0=0$ ainsi que l'égalité $\lambda_{N-k}=\lambda_k$, $k=1, 2, \ldots$ $\ldots, (N-1)/2$.

Les fonctions propres correspondant aux valeurs propres λ_k se déterminent au moyen des formules suivantes:

$$\begin{aligned}\mu_0(i)&=\sqrt{\frac{1}{l}}, && k=0,\\ \mu_k(i)&=\sqrt{\frac{2}{l}}\cos\frac{2k\pi i}{N}, && 1\leqslant k\leqslant\frac{N-1}{2},\\ \mu_k(i)&=\sqrt{\frac{2}{l}}\sin\frac{2(N-k)\pi i}{N}, && \frac{N+1}{2}\leqslant k\leqslant N-1.\end{aligned}\qquad(37)$$

Les fonctions propres (37) sont orthonormées, tandis que les valeurs propres λ_k satisfont aux inégalités $0=\lambda_0<\lambda_k<\lambda_{\frac{N-1}{2}}=$ $=\frac{4}{h^2}\cos^2\frac{\pi}{2N}$, $0<k<N-1$. En outre, toute fonction de maille

$f(i)$ périodique de période N (N étant impair) donnée sur le maillage Ω peut être représentée sous la forme

$$f(i)=\frac{2}{N}\sum_{k=0}^{(N-1)/2}\rho_k\varphi_k\cos\frac{2k\pi i}{N}+\frac{2}{N}\sum_{k=(N+1)/2}^{N-1}\varphi_k\sin\frac{2(N-k)\pi i}{N},$$

où

$$\varphi_k=\sum_{i=0}^{N-1}\rho_k f(i)\cos\frac{2k\pi i}{N},\qquad 0\leqslant k\leqslant\frac{N-1}{2},$$

$$\varphi_k=\sum_{i=0}^{N-1}f(i)\sin\frac{2(N-k)\pi i}{N},\qquad \frac{N+1}{2}\leqslant k\leqslant N-1,$$

ρ_k étant déterminé auparavant.

CHAPITRE II

MÉTHODE DE BALAYAGE

Dans ce chapitre on étudie les différentes variantes de la méthode directe de résolution des équations de mailles, la méthode de balayage. On examine l'application de la méthode à la résolution des équations scalaires et des équations vectorielles.

Au § 1 est construite et étudiée la méthode de balayage pour des équations scalaires triponctuelles. Le § 2 est consacré aux différentes variantes de la méthode de balayage, on y analyse les balayages en flux, cyclique et non monotone. Au § 3 sont étudiées les méthodes des balayages monotone et non monotone appliquées à des équations scalaires pentaponctuelles. Dans le § 4 sont construits les algorithmes du balayage matriciel pour des équations vectorielles bi- et triponctuelles, ainsi que la méthode du balayage orthogonal appliquée à des équations biponctuelles.

§ 1. Méthode du balayage pour les équations triponctuelles

1. Algorithme de la méthode. On a exposé au chapitre I les méthodes de résolution des équations à coefficients constants. Le présent chapitre est dévolu à la construction des méthodes directes de résolution des problèmes aux limites appliquées aux équations aux différences tri- et pentaponctuelles à coefficients variables, ainsi qu'aux équations vectorielles triponctuelles. On y analysera les différentes variantes de la méthode du balayage qui est la méthode d'élimination de Gauss appliquée à des systèmes spéciaux d'équations algébriques linéaires et qui tient compte de la structure en bande de la matrice du système.

Commençons l'étude de la méthode du balayage avec le cas d'équations scalaires. Supposons qu'il s'agit de trouver la solution du système suivant d'équations triponctuelles:

$$
\begin{aligned}
&c_0 y_0 - b_0 y_1 = f_0, && i = 0, \\
&-a_i y_{i-1} + c_i y_i - b_i y_{i+1} = f_i, && 1 \leqslant i \leqslant N-1, \\
&-a_N y_{N-1} + c_N y_N = f_N, && i = N,
\end{aligned}
\tag{1}
$$

ou sous forme vectorielle

$$\mathcal{A} Y = F. \tag{2}$$

où $Y = (y_0, y_1, \ldots, y_N)$ est le vecteur d'inconnues, $F = (f_0, f_1, \ldots \ldots, f_N)$ le vecteur des seconds membres, et $\mathcal{A}$ la matrice carrée $(N+1) \times (N+1)$

$$\mathcal{A} = \begin{Vmatrix} c_0 & -b_0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 \\ -a_1 & c_1 & -b_1 & 0 & \ldots & 0 & 0 & 0 & 0 \\ 0 & -a_2 & c_2 & -b_2 & \ldots & 0 & 0 & 0 & 0 \\ . & . & . & . & \ldots & . & . & . & . \\ 0 & 0 & 0 & 0 & \ldots & -a_{N-2} & c_{N-2} & -b_{N-2} & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & -a_{N-1} & c_{N-1} & -b_{N-1} \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & -a_N & c_N \end{Vmatrix}$$

à coefficients réels ou complexes.

Les systèmes de la forme (1) apparaissent au cas d'approximation triponctuelle des problèmes aux limites sur les équations différentielles ordinaires d'ordre deux à coefficients constants et variables, ainsi que pour la mise en œuvre des schémas aux différences pour équations aux dérivées partielles. Dans ce dernier cas on est souvent obligé de résoudre non pas l'unique problème (1), mais une série de problèmes aux différents seconds membres, le nombre de problèmes dans la série pouvant atteindre plusieurs dizaines et centaines avec un nombre d'inconnues dans chaque problème égal à $N \approx 100$. Aussi faut-il mettre au point des méthodes économiques de résolution des problèmes de la forme (1) dont le nombre d'opérations soit proportionnel à celui d'inconnues. Pour le système (1) la méthode adéquate est la *méthode du balayage*.

La possibilité de construction d'une méthode économique trouve son germe dans la spécificité du système (1). La matrice $\mathcal{A}$ associée à (1) appartient à la classe des matrices raréfiées: des $(N+1)^2$ éléments moins de $3N+1$ éléments seulement sont non nuls. De plus, elle possède une structure en bande (c'est une matrice tridiagonale). Cette disposition régulière des éléments non nuls de la matrice permet d'obtenir des formules de calcul très simples aboutissant à la solution.

Abordons la construction de l'algorithme de résolution du système (1). Rappelons la suite des opérations mises en œuvre dans la méthode d'élimination de Gauss. D'abord, de toutes les équations du système (1) on élimine pour $i = 1, 2, \ldots, N$, à l'aide de la première équation de (1), l'inconnue y_0, ensuite, des équations transformées, pour $i = 2, 3, \ldots, N$, à l'aide de l'équation correspondant à $i = 1$, on élimine l'inconnue y_1, etc. Finalement on obtient une seule équation en y_N. A ce point s'achève la marche en

sens direct de la méthode. Au cours de la marche en sens inverse (par remontée) pour $i = N - 1, N - 2, \ldots, 0$ on obtient y_i au moyen des $y_{i+1}, y_{i+2}, \ldots, y_N$ déjà trouvés et des seconds membres transformés.

En s'inspirant de l'idée de la méthode de Gauss, procédons à l'élimination des inconnues de (1). Introduisons les notations en posant $\alpha_1 = b_0/c_0$, $\beta_1 = f_0/c_0$, et écrivons (1) sous la forme suivante:

$$\begin{aligned} &y_0 - \alpha_1 y_1 = \beta_1, && i = 0, \\ &-a_i y_{i-1} + c_i y_i - b_i y_{i+1} = f_i, && 1 \leqslant i \leqslant N-1, \\ &-a_N y_{N-1} + c_N y_N = f_N, && i = N. \end{aligned} \tag{1'}$$

Prenons les deux premières équations du système (1')

$$y_0 - \alpha_1 y_1 = \beta_1, \quad -a_1 y_0 + c_1 y_1 - b_1 y_2 = f_1.$$

Multiplions la première équation par a_1 et additionnons à la seconde. On aura $(c_1 - a_1\alpha_1)\, y_1 - b_1 y_2 = f_1 + a_1\beta_1$ ou, après division par $c_1 - a_1\alpha_1$,

$$y_1 - \alpha_2 y_2 = \beta_2, \quad \alpha_2 = \frac{b_1}{c_1 - \alpha_1 a_1}, \quad \beta_2 = \frac{f_1 + a_1\beta_1}{c_1 - \alpha_1 a_1}.$$

Toutes les autres équations du système (1') ne contiennent pas y_0, aussi à ce point se termine la première phase des opérations d'élimination. On aboutit ainsi à un nouveau système « raccourci »

$$\begin{aligned} &y_1 - \alpha_2 y_2 = \beta_2, && i = 1, \\ &-a_i y_{i-1} + c_i y_i - b_i y_{i+1} = f_i, && 2 \leqslant i \leqslant N-1, \\ &-a_N y_{N-1} + c_N y_N = f_N, && i = N, \end{aligned} \tag{3}$$

qui ne contient pas l'inconnue y_0 et possède une structure analogue à (1'). Si ce système est résolu, on obtiendra l'inconnue y_0 à l'aide de la formule $y_0 = \alpha_1 y_1 + \beta_1$. Au système (3) on peut appliquer de nouveau le procédé décrit d'élimination des inconnues. A la seconde phase on éliminera l'inconnue y_1, à la troisième l'inconnue y_2, etc. Par suite, à la l-ième phase on obtiendra le système à inconnues $y_l, y_{l+1}, \ldots, y_N$

$$\begin{aligned} &y_l - \alpha_{l+1} y_{l+1} = \beta_{l+1}, && i = l, \\ &-a_i y_{i-1} + c_i y_i - b_i y_{i+1} = f_i, && l+1 \leqslant i \leqslant N-1, \\ &-a_N y_{N-1} + c_N y_N = f_N, && i = N, \end{aligned} \tag{4}$$

et les formules permettant de trouver y_i aux numéros $i \leqslant l - 1$

$$y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1}, \quad i = l-1, l-2, \ldots, 0. \tag{5}$$

Les coefficients α_i et β_i s'obtiennent évidemment au moyen des formules

$$\alpha_{i+1}=\frac{b_i}{c_i-a_i\alpha_i}, \quad \beta_{i+1}=\frac{f_i+a_i\beta_i}{c_i-a_i\alpha_i}, \quad i=1,\ 2,\ \ldots,$$

$$\alpha_1=\frac{b_0}{c_0}, \quad \beta_1=\frac{f_0}{c_0}.$$

En posant dans (4) $l = N - 1$, on obtient le système pour y_N et y_{N-1}

$$y_{N-1}-\alpha_N y_N = \beta_N, \quad -a_N y_{N-1} + c_N y_N = f_N,$$

à partir duquel on trouve $y_N = \beta_{N+1}$, $y_{N-1} = \alpha_N y_N + \beta_N$.

En joignant ces égalités à (5) ($l = N - 1$), on obtient les formules définitives de la recherche des inconnues

$$\begin{aligned} &y_i = \alpha_{i+1}y_{i+1} + \beta_{i+1}, \quad i = N-1,\ N-2,\ \ldots,\ 0, \\ &y_N = \beta_{N+1}, \end{aligned} \tag{6}$$

où α_i et β_i s'obtiennent au moyen des formules de récurrence

$$\alpha_{i+1}=\frac{b_i}{c_i-a_i\alpha_i}, \quad i=1,\ 2,\ \ldots,\ N-1, \quad \alpha_1=\frac{b_0}{c_0}, \tag{7}$$

$$\beta_{i+1}=\frac{f_i+a_i\beta_i}{c_i-a_i\alpha_i}, \quad i=1,\ 2,\ \ldots,\ N, \quad \beta_1=\frac{f_0}{c_0}. \tag{8}$$

Bref, les formules (6)-(8) décrivent la méthode de Gauss qui, appliquée au système (1), a reçu une appellation spéciale, celle de la *méthode du balayage*. Les coefficients α_i et β_i sont appelés *coefficients de balayage*, les formules (7), (8) décrivent la *phase du balayage en sens direct*, la formule (6) la *phase du balayage par remontée*. Vu que les valeurs y_i s'obtiennent ici de proche en proche avec le passage de $i + 1$ à i, les formules (6)-(8) sont quelquefois appelées formules du *balayage à droite*.

Un calcul élémentaire des opérations arithmétiques effectuées dans (6)-(8) montre que la mise en œuvre de la méthode du balayage exige la réalisation de $3N$ multiplications, $2N + 1$ divisions et $3N$ additions et soustractions. Si l'on s'abstient de distinguer les opérations arithmétiques, leur nombre total exigé par la méthode du balayage s'élève à $Q = 8N + 1$. De ce nombre $3N - 2$ opérations sont nécessaires au calcul de α_i et $5N + 3$ opérations à celui de β_i et y_i.

Notons que les coefficients α_i sont indépendants du second membre du système (1) et ne se déterminent que par les coefficients des équations aux différences a_i, b_i, c_i. Aussi s'il s'agit de résoudre une série de problèmes (1) présentant des seconds membres différents, mais possédant une même matrice $\mathcal{A}$, on ne calcule les coefficients de balayage α_i qu'avec la résolution du premier problème de la série. Pour les problèmes suivants, on ne détermine chaque fois que les

coefficients β_i et la solution y_i en utilisant les valeurs de α_i trouvées auparavant. C'est donc la résolution du seul premier problème de la série qui vaut $Q = 8N + 1$ opérations arithmétiques, la résolution de chaque problème suivant n'exigera que $5N + 3$ opérations.

En conclusion, indiquons l'ordre des calculs suivant les formules de la méthode du balayage. En commençant par α_1 et β_1, au moyen des formules (7) et (8), on détermine et on mémorise les coefficients de balayage α_i et β_i. Ensuite, à l'aide des formules (6), on recherche la solution y_i.

2. Méthode des balayages opposés. On a obtenu plus haut les formules du balayage à droite permettant de résoudre le système (1). De façon analogue on déduit les formules du *balayage à gauche*:

$$\xi_i = \frac{a_i}{c_i - b_i\xi_{i+1}}, \qquad i = N-1,\ N-2,\ \ldots,\ 1, \quad \xi_N = \frac{a_N}{c_N}, \tag{9}$$

$$\eta_i = \frac{f_i + b_i\eta_{i+1}}{c_i - b_i\xi_{i+1}}, \qquad i = N-1,\ N-2,\ \ldots,\ 0, \quad \eta_N = \frac{f_N}{c_N}, \tag{10}$$

$$y_{i+1} = \xi_{i+1}y_i + \eta_{i+1}, \quad i = 0,\ 1,\ \ldots,\ N-1, \qquad y_0 = \eta_0. \tag{11}$$

Les valeurs de y_i s'obtiennent ici de proche en proche avec l'accroissement de l'indice i (de gauche à droite).

Il s'avère quelquefois commode de combiner les balayages à droite et à gauche et l'on obtient ainsi la *méthode des balayages opposés*. Il est rationnel d'utiliser cette méthode lorsqu'il s'agit de ne trouver qu'une seule inconnue, par exemple, y_m $(0 \leqslant m \leqslant N)$ ou un groupe d'inconnues se suivant. Cherchons les formules de la méthode des balayages opposés. Soit $1 \leqslant m \leqslant N$ et les formules (7)-(10) donnant $\alpha_1, \alpha_2, \ldots, \alpha_m, \beta_1, \beta_2, \ldots, \beta_m$ et $\xi_N, \xi_{N-1}, \ldots$ $\ldots, \xi_m, \eta_N, \eta_{N-1}, \ldots, \eta_m$. Ecrivons les formules (6), (11) pour la marche par remontée des balayages à droite et à gauche pour $i = m - 1$. On a alors le système

$$y_{m-1} = \alpha_m y_m + \beta_m, \quad y_m = \xi_m y_{m-1} + \eta_m,$$

à partir duquel on déduit y_m:

$$y_m = \frac{\eta_m + \xi_m\beta_m}{1 - \xi_m\alpha_m}.$$

En utilisant y_m trouvé au moyen des formules (6), pour $i =$ $= m-1,\ m-2,\ \ldots,\ 0$, on obtient successivement $y_{m-1}, y_{m-2}, \ldots$ $\ldots, y_0$, et au moyen des formules (11), pour $i = m,\ m+1,\ \ldots$ $\ldots, N$, calculons les $y_{m+1}, y_{m+2}, \ldots, y_N$ restants.

Les formules de la méthode des balayages opposés prennent donc la forme :

$$
\begin{aligned}
\alpha_{i+1} &= \frac{b_i}{c_i - a_i\alpha_i}, & i &= 1, 2, \ldots, m-1, & \alpha_1 &= \frac{b_0}{c_0},\\
\beta_{i+1} &= \frac{f_i + a_i\beta_i}{c_i - a_i\alpha_i}, & i &= 1, 2, \ldots, m-1, & \beta_1 &= \frac{f_0}{c_0},\\
\xi_i &= \frac{a_i}{c_i - b_i\xi_{i+1}}, & i &= N-1, N-2, \ldots, m, & \xi_N &= \frac{a_N}{c_N},\\
\eta_i &= \frac{f_i + b_i\eta_{i+1}}{c_i - b_i\xi_{i+1}}, & i &= N-1, N-2, \ldots, m, & \eta_N &= \frac{f_N}{c_N}
\end{aligned}
\qquad (12)
$$

pour le calcul des coefficients de balayage et

$$
\begin{aligned}
y_i &= \alpha_{i+1}y_{i+1} + \beta_{i+1}, & i &= m-1, m-2, \ldots, 0,\\
y_{i+1} &= \xi_{i+1}y_i + \eta_{i+1}, & i &= m, m+1, \ldots, N-1,\\
y_m &= \frac{\eta_m + \xi_m\beta_m}{1 - \xi_m\alpha_m}
\end{aligned}
\qquad (13)
$$

pour déterminer la solution.

Apparemment, le nombre d'opérations exigé par la recherche de la solution du problème (1) à l'aide de la méthode des balayages opposés est le même qu'au cas des balayages à gauche ou à droite, c'est-à-dire que $Q \approx 8N$. Notons que pour le cas particulier de coefficients constants, $a_i = b_i = 1$, $c_i = c$ pour $i = 1, 2, \ldots$ $\ldots, N-1$ et $b_0 = a_N = 0$, le nombre d'opérations peut être diminué, si N est impair, de la façon suivante. Soit $N = 2M - 1$. Posons dans les formules (12). (13) de la méthode des balayages opposés $m = M$. On a alors $\xi_{N-i+1} = \alpha_i$, $i = 1, 2, \ldots, M$. Par conséquent, le coefficient de balayage ξ_i n'est pas à rechercher et les formules de la méthode des balayages opposés prendront la forme

$$
\begin{aligned}
\alpha_{i+1} &= \frac{1}{c - \alpha_i}, & i &= 1, 2, \ldots, M-1, & \alpha_1 &= 0,\\
\beta_{i+1} &= (f_i + \beta_i)\,\alpha_{i+1}, & i &= 1, 2, \ldots, M-1, & \beta_1 &= \frac{f_0}{c_0},\\
\eta_i &= (f_i + \eta_{i+1})\,\alpha_{N-i+i}, & i &= N-1, N-2, \ldots, M, & \eta_N &= \frac{f_N}{c_N},\\
y_i &= \alpha_{i+1}y_{i+1} + \beta_{i+1}, & i &= M-1, M-2, \ldots, 0,\\
y_{i+1} &= \alpha_{N-i}y_i + \eta_{i+1}, & i &= M, M+1, \ldots, N-1,
\end{aligned}
$$

où $y_M = (\eta_M + \alpha_M\beta_M)/(1 - \alpha_M^2)$.

3. Justification de la méthode du balayage. On a obtenu plus haut les formules de la méthode du balayage sans faire d'hypothèses sur les coefficients du système (1). Arrêtons-nous ici sur la question des exigences imposées à ces coefficients sous le rapport de la pos-

sibilité d'utiliser la méthode à la résolution du problème avec une précision suffisante.

Eclairons la situation. Comme les formules de calcul (6)-(8) de la méthode du balayage contiennent des opérations de division, il faut garantir que le dénominateur $c_i - a_i\alpha_i$ dans (7), (8) ne devienne pas nul. On dira que l'algorithme du balayage à droite est *correct* si $c_i - a_i\alpha_i \neq 0$ pour $i = 1, 2, \ldots, N$. Ensuite, la solution y_i s'obtient suivant la formule de récurrence (6). Cette formule peut engendrer une accumulation d'erreurs d'arrondi associées aux opérations arithmétiques. En effet, supposons que les coefficients de balayage α_i et β_i sont obtenus exacts, mais lors du calcul de y_N est commise une erreur ε_N, c'est-à-dire qu'on a abouti à $\tilde{y}_N = y_N + \varepsilon_N$. La solution $\tilde{y}_i$ étant recherchée suivant la formule (6), $\tilde{y}_i = \alpha_{i+1}\tilde{y}_{i+1} + \beta_{i+1}$, $i = N-1, N-2, \ldots, 0$, l'erreur $\varepsilon_i = \tilde{y}_i - y_i$ vérifiera vraisemblablement l'équation homogène $\varepsilon_i = \alpha_{i+1}\varepsilon_{i+1}$, $i = N-1, N-2, \ldots, 0$, avec ε_N donné. Il en découle que si tous les α_i sont en module supérieurs à l'unité, le résultat se soldera par un fort accroissement de l'erreur ε_0 et, si N est suffisamment grand, la solution réelle $\tilde{y}_i$ différera beaucoup de la solution y_i cherchée.

N'ayant pas la possibilité de nous arrêter en plus de détails sur les questions abordées de la stabilité des calculs de la méthode et du mécanisme de formation de l'instabilité, limitons-nous à la formulation des exigences habituellement imposées à l'algorithme de la méthode du balayage. On exigera que les coefficients de balayage α_i ne soient pas en module supérieurs à l'unité. Cette condition suffisante garantit la non-croissance de l'erreur ε_i dans la situation modèle examinée plus haut. Si la condition $|\alpha_i| \leqslant 1$ est satisfaite, on dira que l'algorithme du balayage à droite est *stable*.

Etablissons les conditions de correction et de stabilité de l'algorithme (6)-(8). Le lemme suivant renferme des conditions suffisantes pour la correction et la stabilité de l'algorithme du balayage à droite.

Lemme 1. *Supposons que les coefficients du système* (1) *sont réels et vérifient les conditions* $|b_0| \geqslant 0$, $|a_N| \geqslant 0$, $|c_0| > 0$, $|c_N| > 0$, $|a_i| > 0$, $|b_i| > 0$, $i = 1, 2, \ldots, N-1$,

$$|c_i| \geqslant |a_i| + |b_i|, \quad i = 1, 2, \ldots, N-1, \tag{14}$$

$$|c_0| \geqslant |b_0|, \quad |c_N| \geqslant |a_N|, \tag{15}$$

en outre dans l'une au moins des inégalités (14) *ou* (15) *est satisfaite l'inégalité stricte, c'est-à-dire que la matrice $\mathcal{A}$ possède une dominance diagonale. Alors pour l'algorithme* (6)-(8) *de la méthode du balayage on a les inégalités $c_i - a_i\alpha_i \neq 0$, $|\alpha_i| \leqslant 1$, $i = 1, 2, \ldots, N$, garantissant la correction et la stabilité de la méthode.*

La démonstration du lemme sera faite par induction. Selon les conditions du lemme et selon (7) il s'ensuit que

$$0 \leqslant |\alpha_1| = \frac{|b_0|}{|c_0|} \leqslant 1. \tag{16}$$

Montrons que de l'inégalité $|\alpha_i| \leqslant 1$ $(i \leqslant N-1)$ et des conditions du lemme s'ensuivent les inégalités

$$c_i - a_i\alpha_i \neq 0, \quad |\alpha_{i+1}| \leqslant 1, \quad i \leqslant N-1. \tag{17}$$

Alors, compte tenu de (16), on obtient qu'il y a lieu aux inégalités $|\alpha_i| \leqslant 1$ pour $i = 1, 2, \ldots, N$ et $c_i - a_i\alpha_i \neq 0$ pour $i = 1, 2, \ldots, N-1$. Pour boucler la démonstration du lemme, il ne reste qu'à prouver l'inégalité $c_N - a_N\alpha_N \neq 0$. Commençons par établir la vérité de (17). Soit $|\alpha_i| \leqslant 1$, $i \leqslant N-1$. Alors de (14)

$$|c_i - a_i\alpha_i| \geqslant |c_i| - |a_i|\,|\alpha_i| \geqslant |b_i| + |a_i|(1 - |\alpha_i|) \geqslant \\ \geqslant |b_i| > 0, \tag{18}$$

et, par suite, $c_i - a_i\alpha_i \neq 0$. Ensuite, de (7) et (18), il vient

$$|\alpha_{i+1}| = \frac{|b_i|}{|c_i - a_i\alpha_i|} \leqslant \frac{|b_i|}{|b_i|} = 1,$$

c'est ce qu'il fallait démontrer.

Il reste à montrer que $c_N - a_N\alpha_N \neq 0$. Pour cela utilisons l'hypothèse suivant laquelle au moins dans l'une des inégalités (14) ou (15) on a une inégalité stricte. Plusieurs cas sont possibles. Si $|c_N| > |a_N|$, en vertu du démontré $|\alpha_N| \leqslant 1$, il s'ensuit que $c_N - a_N\alpha_N \neq 0$. Si l'inégalité stricte est atteinte dans (14) pour un certain i_0, $1 \leqslant i_0 \leqslant N-1$, de (18) on tire que $|c_{i_0} - a_{i_0}\alpha_{i_0}| > |b_{i_0}|$, et, par suite, on a l'inégalité $|\alpha_{i_0+1}| < 1$. Il est aisé d'établir par induction l'inégalité $|\alpha_i| < 1$ pour $i \geqslant i_0 + 1$. Donc dans ce cas on aura $|\alpha_N| < 1$ et, par suite, $c_N - a_N\alpha_N \neq 0$. Si $|c_0| > |b_0|$, l'inégalité $|\alpha_i| < 1$ a lieu à partir de $i = 1$. Et l'on obtient de nouveau $|\alpha_N| < 1$ et $c_N - a_N\alpha_N \neq 0$. Le lemme est démontré.

R e m a r q u e 1. Les conditions de correction et de stabilité de l'algorithme (6)-(8) postulées dans le lemme 1 ne sont que des conditions suffisantes. Ces conditions peuvent être atténuées en permettant à certains des coefficients a_i ou b_i de s'annuler. C'est ainsi, par exemple, si pour un certain $1 \leqslant m \leqslant N-1$ il s'avère que $a_m = 0$, le système (1) se divise en deux systèmes:

$$\begin{aligned} &c_m y_m - b_m y_{m+1} = f_m, && i = m, \\ &-a_i y_{i-1} + c_i y_i - b_i y_{i+1} = f_i, && m+1 \leqslant i \leqslant N-1, \\ &-a_N y_{N-1} + c_N y_N = f_N, && i = N \end{aligned}$$

pour les inconnues $y_m, y_{m+1}, \ldots, y_N$ et

$$\begin{aligned} &c_0y_0 - b_0y_1 = f_0, && i = 0, \\ &-a_iy_{i-1} + c_iy_i - b_iy_{i+1} = f_i, && 1 \leqslant i \leqslant m-2, \\ &-a_{m-1}y_{m-2} + c_{m-1}y_{m-1} = f_{m-1} + b_{m-1}y_m \end{aligned}$$

pour les inconnues $y_0, y_1, \ldots, y_{m-1}$. A chacun de ces systèmes on peut appliquer l'algorithme (6)-(8) si les conditions du lemme 1 sont remplies pour eux. Mais dans ce cas les formules (6)-(8) peuvent être utilisées à la recherche de la solution de tout le système divisé (1) à la fois, l'algorithme restant correct et stable.

Remarque 2. Les conditions du lemme 1 garantissent la correction et la stabilité des algorithmes des balayages à gauche et opposés. Ces conditions sont également requises au cas où le système (1) possède des coefficients a_i, b_i et c_i complexes.

Montrons maintenant qu'avec la satisfaction des conditions du lemme 1 le système (1) possède une solution unique pour tout second membre. En effet, compte tenu du rapport (7), par multiplication directe des matrices on peut montrer que la matrice $\mathcal{A}$ du système (1) se présente sous forme d'un produit de deux matrices triangulaires L et U

$$\mathcal{A} = LU,$$

où

$$L = \left\| \begin{matrix} c_0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 \\ -a_1 & \Delta_1 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 \\ 0 & -a_2 & \Delta_2 & 0 & \ldots & 0 & 0 & 0 & 0 \\ 0 & 0 & -a_3 & \Delta_3 & \ldots & 0 & 0 & 0 & 0 \\ . & . & . & . & \ldots & . & . & . & . \\ 0 & 0 & 0 & 0 & \ldots & \Delta_{N-3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ldots & -a_{N-2} & \Delta_{N-2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & -a_{N-1} & \Delta_{N-1} & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & -a_N & \Delta_N \end{matrix} \right\|,$$

$$U = \left\| \begin{matrix} 1 & -\alpha_1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 1 & -\alpha_2 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 1 & -\alpha_3 & \ldots & 0 & 0 & 0 \\ . & . & . & . & \ldots & . & . & . \\ 0 & 0 & 0 & 0 & \ldots & 1 & -\alpha_{N-1} & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 1 & -\alpha_N \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 1 \end{matrix} \right\|$$

et $\Delta_i = c_i - a_i\alpha_i$, $i = 1, 2, \ldots, N$. Vu que

$$\det \mathcal{A} = \det L \cdot \det U = c_0 \prod_{i=1}^{N} \Delta_i,$$

tandis qu'en vertu du lemme 1 $c_0 \neq 0$ et $\Delta_i \neq 0$ pour $i = 1, 2, \ldots, N$, $\det \mathcal{A} \neq 0$. Aussi le système (1), au cas de satisfaction aux conditions du lemme 1, a-t-il une solution unique, et cette solution peut être obtenue en recourant à la méthode du balayage (6)-(8).

4. Exemples d'application de la méthode du balayage. Passons en revue quelques exemples d'application de la méthode du balayage exposée plus haut.

Exemple 1. *Premier problème aux limites.* Supposons qu'il s'agit de résoudre le problème suivant:

$$\begin{aligned} &(k(x)\, u'(x))' - q(x)\, u(x) = -f(x), \quad 0 < x < l, \\ &u(0) = \mu_1, \quad u(l) = \mu_2, \quad k(x) \geqslant c_1 > 0, \quad q(x) \geqslant 0. \end{aligned} \tag{19}$$

Sur le segment $0 \leqslant x \leqslant l$ construisons un maillage irrégulier quelconque $\bar{\omega} = \{x_i \in [0, l], i = 0, 1, \ldots, N, x_0 = 0, x_N = l\}$ de pas $h_i = x_i - x_{i-1}$, $i = 1, 2, \ldots, N$ et substituons à (19) le problème aux différences suivant:

$$\begin{aligned} &(ay_{\bar{x}})_{\hat{x}, i} - d_i y_i = -\varphi_i, \quad 1 \leqslant i \leqslant N-1, \\ &y_0 = \mu_1, \quad y_N = \mu_2, \end{aligned} \tag{20}$$

où $d_i = q(x_i)$, $\varphi_i = f(x_i)$, tandis que pour a_i utilisons la plus simple des approximations du coefficient $k(x)$: $a_i = k(x_i - 0{,}5h_i)$. En répartissant la différence divisée figurant dans (20) par points

$$(ay_{\bar{x}})_{\hat{x}, i} = \frac{1}{\hbar_i}\left(a_{i+1}\frac{y_{i+1} - y_i}{h_{i+1}} - a_i \frac{y_i - y_{i-1}}{h_i}\right),$$

où $\hbar_i = 0{,}5\,(h_i + h_{i+1})$ est le pas moyen au point x_i, on obtient que le problème (20) s'écrit sous forme de système

$$\begin{aligned} &C_0 y_0 - B_0 y_1 = f_0, && i = 0, \\ &-A_i y_{i-1} + C_i y_i - B_i y_{i+1} = f_i, && 1 \leqslant i \leqslant N-1, \\ &-A_N y_{N-1} + C_N y_N = f_N, && i = N. \end{aligned} \tag{1''}$$

Dans ce cas on a

$$\begin{aligned} &B_0 = A_N = 0, \quad C_0 = C_N = 1, \quad f_0 = \mu_1, \quad f_N = \mu_2, \quad f_i = \varphi_i, \\ &A_i = \frac{a_i}{\hbar_i h_i}, \quad B_i = \frac{a_{i+1}}{\hbar_i h_{i+1}}, \quad C_i = A_i + B_i + d_i, \quad 1 \leqslant i \leqslant N-1. \end{aligned} \tag{21}$$

En vertu du schéma aux différences (20) construit, on a satisfait pour les coefficients a_i et d_i les conditions suivantes: $a_i \geqslant c_1 > 0$, $d_i \geqslant 0$. Aussi, de (21) suit-il que pour (1″) les conditions du lem-

me 1 sont remplies et ce problème peut être résolu par la méthode du balayage.

Exemple 2. *Troisième problème aux limites.* Voyons maintenant les conditions aux limites de troisième espèce:

$$\begin{aligned} &(k(x)\,u'(x))' - q(x)\,u(x) = -f(x), \quad 0 < x < l, \\ &k(0)\,u'(0) = \varkappa_1 u(0) - \mu_1, \\ &-k(l)\,u'(l) = \varkappa_2 u(l) - \mu_2. \end{aligned} \tag{22}$$

Posons satisfaites les conditions suivantes: $k(x) \geqslant c_1 > 0$, $q(x) \geqslant 0$, $\varkappa_1 \geqslant 0$, $\varkappa_2 \geqslant 0$, en outre si $q(x) \equiv 0$, $\varkappa_1^2 + \varkappa_2^2 \neq 0$.

Sur le maillage irrégulier introduit plus haut le problème (22) est approximé par le schéma aux différences suivant:

$$\begin{aligned} &(ay_{\bar{x}})_{\hat{x},\,i} - d_i y_i = -\varphi_i, \quad 1 \leqslant i \leqslant N-1, \\ &\frac{2}{h_1} a_1 y_{x,\,0} = \left(d_0 + \frac{2}{h_1}\varkappa_1\right) y_0 - \left(\varphi_0 + \frac{2}{h_1}\mu_1\right), \quad i = 0, \\ &-\frac{2}{h_N} a_N y_{\bar{x},\,N} = \left(d_N + \frac{2}{h_N}\varkappa_2\right) y_N - \left(\varphi_N + \frac{2}{h_N}\mu_2\right), \quad i = N, \end{aligned} \tag{23}$$

où les coefficients a_i, d_i et φ_i sont choisis de la façon indiquée dans l'exemple 1. En répartissant la différence seconde $(ay_{\bar{x}})_{\hat{x}}$ entre les points, de même que les différences premières

$$y_{x,\,i} = \frac{y_{i+1} - y_i}{h_{i+1}}, \qquad y_{\bar{x},\,i} = \frac{y_i - y_{i-1}}{h_i},$$

ramenons (23) à la forme (1″), où

$$B_0 = \frac{2a_1}{h_1^2}, \quad A_N = \frac{2a_N}{h_N^2}, \quad C_0 = B_0 + d_0 + \frac{2}{h_1}\varkappa_1,$$

$$C_N = A_N + d_N + \frac{2}{h_N}\varkappa_2, \quad f_0 = \varphi_0 + \frac{2}{h_1}\mu_1, \quad f_N = \varphi_N + \frac{2}{h_N}\mu_2,$$

$$A_i = \frac{a_i}{\hbar_i h_i}, \quad B_i = \frac{a_{i+1}}{\hbar_i h_{i+1}}, \ C_i = A_i + B_i + d_i, \ f_i = \varphi_i, \ 1 \leqslant i \leqslant N-1.$$

Il est aisé de vérifier que dans ce cas les conditions du lemme 1 sont également vérifiées.

Exemple 3. *Schémas aux différences pour l'équation de conductibilité thermique.* Examinons le premier problème aux limites pour le cas de l'équation de conductibilité thermique:

$$\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < l, \quad t > 0, \\ &u(0,\ t) = \mu_1(t), \quad u(l,\ t) = \mu_2(t), \\ &u(x,\ 0) = u_0(x). \end{aligned} \tag{24}$$

Introduisons sur le plan $(x,\ t)$ le maillage $\bar{\omega} = \{(x_i,\ t_n),\ x_i = ih,\ i = 0, 1, \ldots, N,\ h = l/N,\ t_n = n\tau,\ n = 0, 1 \ldots\}$ de pas h

dans l'espace et τ par rapport au temps. Approximons (24) par le schéma aux différences

$$y_{t,i} = \sigma y_{\bar{x}x,i}^{n+1} + (1-\sigma) y_{\bar{x}x,i}^{n}, \quad 1 \leqslant i \leqslant N-1,$$
$$y_0^n = \mu_1(t_n), \quad y_N^n = \mu_2(t_n), \quad y_i^0 = u_0(x_i), \quad n = 0, 1, \ldots, \tag{25}$$

où σ est le paramètre réel, $y_i^n = y(x_i, t_n)$,

$$y_{\bar{x}x,i} = \frac{1}{h^2}(y_{i+1} - 2y_i + y_{i-1}), \quad y_{t,i} = \frac{1}{\tau}(y_i^{n+1} - y_i^n). \tag{26}$$

Il est connu (voir, par exemple, [9]), que le schéma (25) possède l'approximation $O(\tau + h^2)$ pour tout σ, $O(\tau^2 + h^2)$ pour $\sigma = 0{,}5$ et l'approximation $O(\tau^2 + h^4)$ pour $\sigma = 1/2 - h^2/(12\tau)$. La condition de stabilité du schéma (25) suivant les données initiales a la forme

$$\sigma \geqslant 1/2 - h^2/(4\tau). \tag{27}$$

Appliquons maintenant la méthode de résolution des équations (25) en faisant intervenir y_i^{n+1}. En posant y_i^n déjà connu, écrivons (25) sous la forme suivante:

$$\frac{1}{\sigma\tau} y_i^{n+1} - y_{\bar{x}x,i}^{n+1} = \varphi_i^n, \quad 1 \leqslant i \leqslant N-1,$$
$$y_0^{n+1} = \mu_1(t_{n+1}), \quad y_N^{n+1} = \mu_2(t_{n+1}),$$

où $\varphi_i^n = \frac{1}{\sigma\tau} y_i^n + \left(\frac{1}{\sigma} - 1\right) y_{\bar{x}x,i}^n$ si $\sigma \neq 0$. En profitant de (26), ramenons ce schéma à la forme (1″), où $B_0 = A_N = 0$, $C_0 = C_N = 1$, $f_0 = \mu_1(t_{n+1})$, $f_N = \mu_2(t_{n+1})$, $A_i = B_i = \frac{1}{h^2}$, $C_i = A_i + B_i + \frac{1}{\sigma\tau}$, $f_i = \varphi_i^n$, $1 \leqslant i \leqslant N-1$. Cherchons les conditions pour lesquelles le système construit (1″) peut être résolu par la méthode du balayage. Il suit du lemme 1 que pour ce faire il faut remplir la condition $|2/h^2 + 1/(\sigma\tau)| \geqslant 2/h^2$. En résolvant cette inégalité, on obtient la condition suffisante de l'applicabilité de la méthode du balayage $\sigma \geqslant -h^2/(4\tau)$. En comparant cette inégalité à (27), on obtient que si pour le schéma (25) est vérifiée la condition de stabilité (27), on peut, pour la recherche de la solution sur la couche supérieure, utiliser la méthode du balayage.

E x e m p l e 4. *Equation non stationnaire de Schrödinger.* Examinons l'équation non stationnaire de Schrödinger $i\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$, $0 < x < l$, $t > 0$, $u(0, t) = u(l, t) = 0$, $u(0, x) = u_0(x)$, $i = \sqrt{-1}$.

Cette équation, de même que l'équation de la conductibilité thermique (24), admet la construction d'un schéma à deux couches

aux poids

$$\begin{gathered} iy_{t,k} = \sigma y^{n+1}_{\bar{x}x,k} + (1-\sigma)\, y^{n}_{\bar{x}x,k}, \qquad 1 \leqslant k \leqslant N-1, \\ y^n_0 = y^n_N = 0, \qquad y^0_k = u_0(x_k), \end{gathered} \tag{28}$$

où le paramètre $\sigma = \sigma_0 + i\sigma_1$ peut prendre des valeurs dans le plan complexe. Le schéma (28) possède une erreur d'approximation $O(\tau + h^2)$ pour tout σ, pour $\sigma = 0{,}5$ elle est égale à $O(\tau^2 + h^2)$ et pour $\sigma = 1/2 - h^2 i/(12\tau)$ l'erreur d'approximation vaut $O(\tau^2 + h^4)$. La condition de stabilité d'après les données initiales est de la forme

$$\sigma_0 = \operatorname{Re} \sigma \geqslant 0{,}5. \tag{29}$$

Le schéma (28) se réduit généralement au système (1″), les conditions du lemme 1 prenant la forme suivante : $|\, 2/h^2 + i/(\sigma\tau)\, | \geqslant 2/h^2$. En résolvant cette inégalité, on aboutit à ce que la solution du schéma (28), obtenue sur la couche supérieure par la méthode du balayage, est correcte à condition que $\sigma_1 = \operatorname{Im} \sigma \geqslant -h^2/(4\tau)$.

Par conséquent, dans l'exemple considéré la condition de l'applicabilité de la méthode du balayage ne coïncide pas avec celle de la stabilité du schéma aux différences relativement aux données initiales.

§ 2. Variantes de la méthode du balayage

1. Méthode du balayage en flux. Examinons la variante de la méthode du balayage utilisée à la résolution des problèmes de différences à coefficients fortement variables. Les exemples de ces problèmes peuvent être empruntés à l'hydrodynamique de la conductibilité thermique et à l'hydrodynamique magnétique, où les coefficients de conductibilités thermique et électrique sont fonction des paramètres thermodynamiques du milieu. Au cas de problèmes thermiques on peut se heurter à des plages adiabatiques à conductibilité thermique nulle, ainsi qu'à des plages isothermiques, où la conductibilité thermique est infiniment grande. Dans les problèmes magnétiques, on rencontre respectivement des bandes à conductibilité parfaite et des bandes de non-conductibilité électrique.

Souvent dans ces problèmes il s'agit, en plus de la solution, de trouver le flux de chaleur (problème thermique). L'opération de résolution des équations aux différences du second ordre au moyen de formules ordinaires de la méthode du balayage, auxquelles se réduisent les schémas aux différences de ces problèmes, se solde par une perte importante en précision. L'utilisation subséquente de la dérivation numérique pour le calcul du flux de chaleur conduit à un résultat inacceptable. On arrive à parer à ce défaut en recourant à la méthode dite *méthode du balayage en flux*. Les formules de cette

méthode du balayage peuvent être obtenues par transformation des formules du balayage ordinaire.

Voyons donc le problème de différences aux limites

$$\begin{gathered} -a_i y_{i-1} + c_i y_i - a_{i+1} y_{i+1} = f_i, \quad 1 \leqslant i \leqslant N-1, \\ y_0 - \varkappa_1 y_1 = \mu_1, \quad y_N - \varkappa_2 y_{N-1} = \mu_2, \end{gathered} \tag{1}$$

où

$$c_i = a_i + a_{i+1} + d_i, \quad 0 < a_i < \infty, \tag{2}$$

$$d_i > 0, \quad i = 1, 2, \ldots, N-1, \quad |\varkappa_1| \leqslant 1, \quad |\varkappa_2| \leqslant 1. \tag{3}$$

Les formules du balayage à droite (voir (6)-(8), § 1) du problème (1), compte tenu de (2), prennent la forme

$$y_i = \bar{\alpha}_{i+1} y_{i+1} + \bar{\beta}_{i+1}, \quad i = N-1, N-2, \ldots, 0, \quad y_N = \frac{\mu_2 + \varkappa_2 \bar{\beta}_N}{1 - \varkappa_2 \bar{\alpha}_N}, \tag{4}$$

$$\bar{\alpha}_{i+1} = \frac{a_{i+1}}{a_{i+1} + d_i + a_i (1 - \bar{\alpha}_i)}, \quad i = 1, 2, \ldots, N-1, \quad \bar{\alpha}_1 = \varkappa_1, \tag{5}$$

$$\bar{\beta}_{i+1} = (f_i + a_i \bar{\beta}_i) \frac{\bar{\alpha}_{i+1}}{a_{i+1}}, \quad i = 1, 2, \ldots, N-1, \quad \bar{\beta}_1 = \mu_1. \tag{6}$$

Introduisons une nouvelle fonction de maille inconnue (flux) suivant la formule

$$w_i = -a_i (y_i - y_{i-1}), \quad i = 1, 2, \ldots, N, \tag{7}$$

et récrivons (1) sous la forme

$$\begin{aligned} & w_{i+1} - w_i + d_i y_i = f_i, && 1 \leqslant i \leqslant N-1, \\ & y_0 - \varkappa_1 y_1 = \mu_1, && i = 0, \\ & -\varkappa_2 w_N + a_N (1 - \varkappa_2) y_N = a_N \mu_2, && i = N. \end{aligned} \tag{8}$$

De (7), il vient

$$y_i = y_{i+1} + \frac{1}{a_{i+1}} w_{i+1}, \quad i = 0, 1, \ldots, N-1,$$

et portons cette expression dans (4). On trouve ainsi la relation liant y_{i+1} et w_{i+1}:

$$w_{i+1} + a_{i+1} (1 - \bar{\alpha}_{i+1}) y_{i+1} = a_{i+1} \bar{\beta}_{i+1}, \quad i = 0, 1, \ldots, N-1.$$

En introduisant les notations

$$\alpha_i = a_i (1 - \bar{\alpha}_i), \quad \beta_i = \alpha_i \bar{\beta}_i, \quad i = 1, 2, \ldots, N,$$

récrivons cette relation sous la forme

$$w_i + \alpha_i y_i = \beta_i, \quad i = 1, 2, \ldots, N. \tag{9}$$

Remarquons que les équations (8), (9) forment un système algébrique comprenant $2N+1$ équations en $2N+1$ inconnues $y_0, y_1, \ldots, y_N$ et $w_1, w_2, \ldots, w_N$. La structure de ce système est telle qu'il se scinde en deux systèmes indépendants pour les inconnues $y_0, y_1, \ldots, y_N$ et $w_1, w_2, \ldots, w_N$. Construisons ces systèmes.

A partir de (9) exprimons y_i: $y_i = (\beta_i - w_i)/\alpha_i$, $i = 1, 2, \ldots, N$ et portons dans les équations du système (8) pour $i = 1, 2, \ldots, N$. On obtient ainsi les équations

$$\begin{aligned} w_i &= \frac{\alpha_i}{\alpha_i+d_i} w_{i+1} + \frac{d_i\beta_i - \alpha_i f_i}{\alpha_i+d_i}, \qquad i = N-1,\ N-2, \ldots, 1, \\ w_N &= \frac{a_N [(1-\varkappa_2)\beta_N - \alpha_N\mu_2]}{(1-\varkappa_2)a_N + \alpha_N\varkappa_2}, \end{aligned} \tag{10}$$

qui, une fois résolues, fournissent successivement tous les w_i.

Obtenons maintenant les équations pour y_i. Pour cela exprimons w_i sur la base de (9): $w_i = -\alpha_i y_i + \beta_i$, $i = 1, 2, \ldots, N$, et portons dans (8) pour $i = 1, 2, \ldots, N$. On aboutit ainsi aux équations

$$\begin{aligned} y_i &= \frac{\alpha_{i+1}}{\alpha_i+d_i} y_{i+1} + \frac{f_i - \beta_{i+1} + \beta_i}{\alpha_i+d_i}, \qquad i = N-1,\ N-2, \ldots, 1, \\ y_0 &= \varkappa_1 y_1 + \mu_1, \\ y_N &= \frac{\varkappa_2\beta_N + a_N\mu_2}{(1-\varkappa_2)a_N + \alpha_N\varkappa_2} \end{aligned} \tag{11}$$

servant successivement au calcul de y_i.

Ecrivons les formules de récurrence permettant de déterminer α_i et β_i. Sur la base de (5) et (6), on trouve

$$\begin{gathered} \alpha_{i+1} = a_{i+1}(1-\bar{\alpha}_{i+1}) = \frac{a_{i+1}[a_i(1-\bar{\alpha}_i)+d_i]}{a_{i+1}+d_i+a_i(1-\bar{\alpha}_i)} = \frac{a_{i+1}(\alpha_i+d_i)}{a_{i+1}+\alpha_i+d_i}, \\ i = 1, 2, \ldots, N-1, \quad \alpha_1 = a_1(1-\varkappa_1), \end{gathered} \tag{12}$$

$$\beta_{i+1} = a_{i+1}\bar{\beta}_{i+1} = \frac{a_{i+1}(f_i+\beta_i)}{a_{i+1}+\alpha_i+d_i}, \qquad i = 1, 2, \ldots, N-1, \quad \beta_1 = a_1\mu_1. \tag{13}$$

Des conditions (2), (3) et des formules (12) il suit que $\alpha_i \geqslant 0$. Dans ce cas le coefficient $\alpha_i/(\alpha_i + d_i)$ dans la formule (10) ne dépasse pas l'unité, ce qui garantit la stabilité de l'algorithme lors du calcul de w_i. Ensuite, puisque des conditions $\alpha_i \geqslant 0$ et $d_i > 0$ il suit que $a_{i+1} < a_{i+1} + \alpha_i + d_i$, en vertu de (12) l'inégalité $\alpha_{i+1} < \alpha_i + d_i$ se vérifie. Aussi le coefficient $\alpha_{i+1}/(\alpha_i + d_i)$ de la formule (11) est-il toujours inférieur à l'unité, ce qui garantit la stabilité lors du calcul de y_i. Notons que le dénominateur dans les expressions de w_N et y_N est toujours supérieur à zéro.

Bref, l'algorithme de la méthode de balayage en flux se décrit à l'aide des formules (10)-(13). Notons qu'il est rationnel de se

servir des formules de récurrence mentionnées de α_i et β_i ainsi que des expressions de y_N et w_N dans le cas où $a_{i+1} < 1$. Si $a_{i+1} \geqslant 1$, il est recommandé d'utiliser les formules suivantes qu'on obtient à partir de (10)-(13) en divisant le numérateur et le dénominateur des fractions par a_{i+1}:

$$\alpha_{i+1} = \frac{\alpha_i + d_i}{1 + (\alpha_i + d_i)/a_{i+1}}, \qquad \beta_{i+1} = \frac{f_i + \beta_i}{1 + (\alpha_i + d_i)/a_{i+1}},$$

$$y_N = \frac{\varkappa_2 \beta_N / a_N + \mu_2}{1 - \varkappa_2 + \varkappa_2 \alpha_N / a_N}, \qquad w_N = \frac{(1 - \varkappa_2)\beta_N - \alpha_N \mu_2}{1 - \varkappa_2 + \varkappa_2 \alpha_N / a_N}.$$

Calculons le nombre d'opérations arithmétiques qu'il est nécessaire d'effectuer pour résoudre (10)-(13). Avec une organisation rationnelle des calculs, quand les expressions communes à plusieurs formules ne se calculent qu'une fois et les facteurs communs à plusieurs termes additionnés sont sortis des parenthèses, le nombre d'opérations impliquées dans (10)-(13) est égal à $Q = 21N + 1$. Ce nombre dépasse de deux fois celui dépensé pour obtenir, en se servant des formules du balayage ordinaire, la solution y_i du problème (1) et, ensuite, à l'aide de la formule (7), le flux w_i.

2. Méthode du balayage cyclique. Voyons maintenant le système suivant:

$$-a_i y_{i-1} + c_i y_i - b_i y_{i+1} = f_i, \quad i = 0, \pm 1, \pm 2, \ldots, \tag{14}$$

dont les coefficients et le second membre sont périodiques de période N:

$$a_i = a_{i+N}, \quad b_i = b_{i+N}, \quad c_i = c_{i+N}, \quad f_i = f_{i+N}. \tag{15}$$

On aboutit aux systèmes du type (14), (15) lors de l'étude, par exemple, des schémas aux différences triponctuels promus à la recherche des solutions périodiques d'équations différentielles ordinaires du second ordre, ainsi qu'à la recherche de la solution approchée des équations à dérivées partielles en coordonnées cylindriques et sphériques.

Avec l'accomplissement des conditions (15), la solution du système (14), si cette dernière existe, sera aussi périodique de période N, c'est-à-dire

$$y_i = y_{i+N}. \tag{16}$$

Aussi suffit-il de trouver la solution y_i, par exemple, pour $i = 0, 1, \ldots, N-1$. On peut alors écrire le problème (14)-(16) de la façon suivante:

$$\begin{aligned} -a_0 y_{N-1} + c_0 y_0 - b_0 y_1 &= f_0, \qquad i = 0, \\ -a_i y_{i-1} + c_i y_i - b_i y_{i+1} &= f_i, \qquad 1 \leqslant i \leqslant N-1, \end{aligned} \tag{17}$$

$$y_N = y_0. \tag{18}$$

La condition (18) a été ajoutée au système (17) pour ne pas exclure

y_N de l'équation du système pour $i = N - 1$ en lui substituant y_0. Cela permet de conserver une forme unique aux équations (17) pour $i = 1, 2, \ldots, N - 1$.

Si l'on introduit les vecteurs d'inconnues $Y = (y_0, y_1, \ldots, y_{N-1})$ et du second membre $F = (f_0, f_1, \ldots, f_{N-1})$, (17) et (18) peuvent être écrits sous forme vectorielle $\mathcal{A}Y = F$, où

$$\mathcal{A} = \begin{Vmatrix} c_0 & -b_0 & 0 & 0 & \ldots & 0 & 0 & -a_0 \\ -a_1 & c_1 & -b_1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & -a_2 & c_2 & -b_2 & \ldots & 0 & 0 & 0 \\ . & . & . & . & \ldots & . & . & . \\ 0 & 0 & 0 & 0 & \ldots & c_{N-3} & -b_{N-3} & 0 \\ 0 & 0 & 0 & 0 & \ldots & -a_{N-2} & c_{N-2} & -b_{N-2} \\ -b_{N-1} & 0 & 0 & 0 & \ldots & 0 & -a_{N-1} & c_{N-1} \end{Vmatrix}$$

est la matrice du système (17), (18). La présence de coefficients a_0 et b_{N-1} différents de zéro dans (17) ne permet pas de résoudre ce système par la méthode du balayage décrite dans le § 1. Pour la recherche de la solution du système (17), (18), construisons la variante de la méthode du balayage qu'on appelle *méthode du balayage cyclique*.

La solution du problème (17), (18) sera recherchée sous forme d'une combinaison linéaire des fonctions de mailles u_i et v_i

$$y_i = u_i + y_0 v_i, \quad 0 \leqslant i \leqslant N, \tag{19}$$

où u_i est la solution du problème aux limites triponctuel inhomogène

$$\begin{gathered} -a_i u_{i-1} + c_i u_i - b_i u_{i+1} = f_i, \quad 1 \leqslant i \leqslant N-1, \\ u_0 = 0, \quad u_N = 0 \end{gathered} \tag{20}$$

aux conditions aux limites homogènes, tandis que v_i est la solution du problème aux limites triponctuel homogène

$$\begin{gathered} -a_i v_{i-1} + c_i v_i - b_i v_{i+1} = 0, \quad 1 \leqslant i \leqslant N-1, \\ v_0 = 1, \quad v_N = 1 \end{gathered} \tag{21}$$

aux conditions aux limites inhomogènes.

Voyons à quelle condition y_i de (19) est la solution cherchée. En multipliant (21) par y_0, en additionnant à (20) et compte tenu de (19), on constate que pour $i = 1, 2, \ldots, N - 1$ les équations du système (17) sont vérifiées. Des conditions aux limites pour u_i et v_i il suit que la relation (18) est satisfaite. Donc, si y_i, déterminé au moyen de la formule (19), satisfait, pour $i = 0$, l'équation du système (17), restée inutilisée, le problème est résolu. Portant (19) dans cette équation, on obtient

$$-a_0 u_{N-1} - a_0 y_0 v_{N-1} + c_0 y_0 - b_0 u_1 - b_0 y_0 v_1 = f_0. \tag{22}$$

Donc si l'on choisit y_0 à l'aide de la formule

$$y_0 = \frac{f_0 + a_0 u_{N-1} + b_0 u_1}{c_0 - a_0 v_{N-1} - b_0 v_1}, \tag{23}$$

l'égalité (22) sera vérifiée et, par suite, la solution du problème (17), (18) peut être obtenue suivant la formule (19).

Fixons maintenant l'attention sur la solution des systèmes (20) et (21). Ils sont des cas particuliers des systèmes d'équations triponctuels pour lesquels a été construite au § 1 la méthode du balayage. Pour (20) et (21), les formules du balayage prennent la forme suivante:

$$u_i = \alpha_{i+1} u_{i+1} + \beta_{i+1}, \quad i = N-1, N-2, \ldots, 1, \quad u_N = 0,$$
$$v_i = \alpha_{i+1} v_{i+1} + \gamma_{i+1}, \quad i = N-1, N-2, \ldots, 1, \quad v_N = 1, \tag{24}$$

où les coefficients de balayage α_i, β_i et γ_i s'obtiennent au moyen des formules suivantes:

$$\alpha_{i+1} = \frac{b_i}{c_i - a_i \alpha_i}, \qquad i = 1, 2, \ldots, N, \qquad \alpha_1 = 0, \tag{25}$$

$$\beta_{i+1} = \frac{f_i + a_i \beta_i}{c_i - a_i \alpha_i}, \qquad i = 1, 2, \ldots, N, \qquad \beta_1 = 0, \tag{26}$$

$$\gamma_{i+1} = \frac{a_i \gamma_i}{c_i - a_i \alpha_i}, \qquad i = 1, 2, \ldots, N, \qquad \gamma_1 = 1. \tag{27}$$

Transformons (23). De (24) on tire $u_{N-1} = \alpha_N u_N + \beta_N = \beta_N$, $v_{N-1} = \gamma_N + \alpha_N$. Portons ces expressions dans (23) et tenons compte des conditions (15), (25)-(27):

$$y_0 = \frac{f_N + a_N \beta_N + \beta_N u_1}{c_N - a_N \alpha_N - a_N \gamma_N - b_N v_1} = \frac{\beta_{N+1} + \alpha_{N+1} u_1}{1 - \gamma_{N+1} - \alpha_{N+1} v_1}.$$

On a construit l'algorithme de la solution du problème (17), (18) qui porte le nom de la méthode du balayage cyclique:

$$\alpha_2 = b_1/c_1, \quad \beta_2 = f_1/c_1, \quad \gamma_2 = a_1/c_1,$$
$$\alpha_{i+1} = \frac{b_i}{c_i - a_i \alpha_i}, \quad \beta_{i+1} = \frac{f_i + a_i \beta_i}{c_i - a_i \alpha_i}, \quad \gamma_{i+1} = \frac{a_i \gamma_i}{c_i - a_i \alpha_i},$$
$$i = 2, 3, \ldots, N;$$
$$u_{N-1} = \beta_N, \quad v_{N-1} = \alpha_N + \gamma_N, \tag{28}$$
$$u_i = \alpha_{i+1} u_{i+1} + \beta_{i+1}, \quad v_i = \alpha_{i+1} v_{i+1} + \gamma_{i+1},$$
$$i = N-2, N-3, \ldots, 1;$$
$$y_0 = \frac{\beta_{N+1} + \alpha_{N+1} u_1}{1 - \gamma_{N+1} - \alpha_{N+1} v_1}, \quad y_i = u_i + y_0 v_i, \quad i = 1, 2, \ldots, N-1.$$

Un calcul élémentaire montre que pour sa mise en œuvre il faut effectuer $6(N-1)$ multiplications, $5N-3$ additions et soustractions et $3N+1$ divisions. Si l'on ne distingue pas entre les opérations arithmétiques, leur nombre total s'élève à $Q = 14N - 8$ opérations.

Etudions la question de l'applicabilité et de la stabilité de l'algorithme (28). On a le

Lemme 2. *Soient les coefficients du système* (14), (15) *satisfaisant aux conditions*

$$|a_i| > 0, \quad |b_i| > 0, \quad |c_i| \geqslant |a_i| + |b_i|, \quad i = 1, 2, \ldots, N, \tag{29}$$

et que, de plus, on ait $1 \leqslant i_0 \leqslant N$ *pour lequel* $|c_{i_0}| > |a_{i_0}| + |b_{i_0}|$. *Dans ce cas*

$$c_i - a_i\alpha_i \neq 0, \quad |\alpha_i| \leqslant 1, \quad |\alpha_i| + |\gamma_i| \leqslant 1, \quad i = 2, 3, \ldots, N,$$
$$1 - \gamma_{N+1} - \alpha_{N+1}v_1 \neq 0.$$

En effet, comme α_i, β_i et γ_i sont des coefficients de balayage de la méthode de balayage à droite utilisée à la résolution des problèmes (20) et (21), tandis qu'en vertu de (29) les conditions du lemme 1 sont remplies, il s'ensuit du lemme 1 que les inégalités

$$\begin{gathered} c_i - a_i\alpha_i \neq 0, \quad |\alpha_i| \leqslant 1, \quad i = 2, 3, \ldots, N, \\ |c_i - a_i\alpha_i| \geqslant |c_i| - |a_i|\,|\alpha_i| \geqslant |b_i| > 0 \end{gathered} \tag{30}$$

sont vraies.

Ensuite, sur la base des conditions du lemme 2, $|a_1| + |b_1| \leqslant$ $\leqslant |c_1|$ et, partant, $|\alpha_2| + |\gamma_2| \leqslant 1$. De là, par induction, on obtient les inégalités

$$|\alpha_i| + |\gamma_i| \leqslant 1, \quad i = 2, 3, \ldots, N, \tag{31}$$

car

$$|\alpha_{i+1}| + |\gamma_{i+1}| = \frac{|b_i| + |a_i|\,|\gamma_i|}{|c_i - a_i\alpha_i|} \leqslant \frac{|a_i| + |b_i| - |a_i|\,(1 - |\gamma_i|)}{|c_i| - |a_i|\,|\alpha_i|} \leqslant$$
$$\leqslant \frac{|a_i| + |b_i| - |a_i|\,|\alpha_i|}{|c_i| - |a_i|\,|\alpha_i|} \leqslant 1$$

et on aboutit donc à (30). Notons que $|c_i| > |a_i| + |b_i|$ pour $i = i_0$ et, partant, $|\alpha_{i_0+1}| + |\gamma_{i_0+1}| < 1$. Il s'ensuit que pour $i \geqslant i_0 + 1$ on a l'inégalité stricte $|\alpha_i| + |\gamma_i| < 1$. Vu que $1 \leqslant i_0 \leqslant N$, $|\alpha_{N+1}| + |\gamma_{N+1}| < 1$.

Il nous reste à montrer que $1 - \gamma_{N+1} - \alpha_{N+1}v_1 \neq 0$. Sur la base de (28) et (31), il vient

$$|v_{N-1}| \leqslant |\alpha_N| + |\gamma_N| \leqslant 1,$$

ensuite, par induction, démontrons les inégalités $|v_i| \leqslant 1$, $1 \leqslant i \leqslant N-1$, vu qu'en vertu de (31)

$$|v_i| \leqslant |\alpha_{i+1}|\,|v_{i+1}| + |\gamma_{i+1}| \leqslant |\alpha_{i+1}| + |\gamma_{i+1}| \leqslant 1.$$

En particulier, $|v_1| \leqslant 1$. De là, compte tenu de l'inégalité démontrée $|\alpha_{N+1}| + |\gamma_{N+1}| < 1$, on tire que

$$|1 - \gamma_{N+1} - \alpha_{N+1}v_1| \geqslant 1 - |\gamma_{N+1}| - |\alpha_{N+1}|\,|v_1| \geqslant$$
$$\geqslant 1 - |\alpha_{N+1}| - |\gamma_{N+1}| > 0.$$

Le lemme 2 est complètement démontré.

En conclusion, remarquons que du second membre f_i dépend le coefficient de balayage β_i et, partant, u_i et y_i. Les coefficients de balayage α_i et γ_i, de même que v_i, ne dépendent pas de f_i et, lors de la résolution du seul premier problème de la série, ils sont calculés et retenus. Cela permet de résoudre le second problème, de même que chaque problème suivant de la série, en $Q = 9N - 4$ opérations.

3. Méthode du balayage pour des systèmes complexes. Continuons à construire les variantes de la méthode du balayage servant à la résolution de systèmes d'équations aux différences dont les matrices ne sont pas tridiagonales. Au point 2 la méthode du balayage cyclique était utilisée à la résolution de systèmes dont les matrices ne contenaient en dehors des principales diagonales que deux éléments non nuls. Voyons maintenant un cas plus général.

Supposons qu'il s'agit de résoudre le système d'équations suivant:

$$\begin{aligned} &c_0y_0 - \sum_{j=1}^{N-1} d_jy_j - \psi_0y_N = f_0, && i=0,\\ &-\varphi_iy_0 - a_iy_{i-1} + c_iy_i - b_iy_{i+1} - \psi_iy_N = f_i, && 1 \leqslant i \leqslant N-1,\\ &-\varphi_Ny_0 - \sum_{j=1}^{N-1} g_jy_j + c_Ny_N = f_N, && i=N. \end{aligned} \qquad (32)$$

Le système de la forme (32) apparaît lors de l'approximation des équations différentielles ordinaires du second ordre aux cas de conditions aux limites liées, de la recherche de solutions satisfaisant aux conditions complémentaires du type d'intégrales, ainsi que dans une série d'autres cas. En particulier, on peut écrire sous cette forme tous les systèmes d'équations aux différences étudiés plus haut. Par exemple, si l'on pose dans (32)

$$d_1 = b_0,\quad d_{N-1} = a_0,\quad d_i = 0,\quad 2 \leqslant i \leqslant N-2,$$
$$\varphi_i = \psi_i = g_i = 0,\quad 1 \leqslant i \leqslant N-1,$$
$$\psi_0 = 0,\quad \varphi_N = c_N = 1,\quad f_N = 0,$$

on obtient le problème (17), (18).

Si l'on introduit les vecteurs $Y = (y_0, y_1, \ldots, y_N)$ et $F = (f_0, \ldots, f_N)$, 32) peut s'écrire sous forme vectorielle $\mathcal{A}Y = F$, où

$$\mathcal{A} = \left\| \begin{array}{c|ccccccc|c} c_0 & -d_1 & -d_2 & -d_3 & \ldots & -d_{N-3} & -d_{N-2} & -d_{N-1} & -\psi_0 \\ \hline -\varphi_1 - a_1 & c_1 & -b_1 & 0 & \ldots & 0 & 0 & 0 & -\psi_1 \\ -\varphi_2 & -a_2 & c_2 & -b_2 & \ldots & 0 & 0 & 0 & -\psi_2 \\ -\varphi_3 & 0 & -a_3 & c_3 & \ldots & 0 & 0 & 0 & -\psi_3 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -\varphi_{N-3} & 0 & 0 & 0 & \ldots & c_{N-3} & -b_{N-3} & 0 & -\psi_{N-3} \\ -\varphi_{N-2} & 0 & 0 & 0 & \ldots & -a_{N-2} & c_{N-2} & -b_{N-2} & -\psi_{N-2} \\ -\varphi_{N-1} & 0 & 0 & 0 & \ldots & 0 & -a_{N-1} & c_{N-1} & -b_{N-1} - \psi_{N-1} \\ \hline -\varphi_N & -g_1 & -g_2 & -g_3 & \ldots & -g_{N-3} & -g_{N-2} & -g_{N-1} & c_N \end{array} \right\| .$$

On voit que la matrice $\mathcal{A}$ est obtenue en bordant la matrice tridiagonale de colonnes et de lignes sur tous les quatre côtés. Notons qu'avec une autre mise en ordre des inconnues $Y^* = (y_1, y_2, \ldots, y_N, y_0)$ le système (32) s'écrira sous la forme $\mathcal{A}^* Y^* = F^*$, dont la matrice $\mathcal{A}^*$ s'obtient par bordage de la même matrice tridiagonale, mais cette fois seulement avec deux colonnes à droite et deux lignes en bas.

Passons à l'élaboration de la méthode de résolution du problème (32). La solution du problème (32) sera recherchée sous forme d'une combinaison linéaire de trois fonctions de mailles u_i, v_i et w_i:

$$y_i = u_i + y_0 v_i + y_N w_i, \quad 0 \leqslant i \leqslant N, \tag{33}$$

où u_i, v_i et w_i sont les solutions des problèmes aux limites triponctuels suivants:

$$\left. \begin{array}{l} -a_i u_{i-1} + c_i u_i - b_i u_{i+1} = f_i, \quad 1 \leqslant i \leqslant N-1, \\ u_0 = 0, \quad u_N = 0; \end{array} \right\} \tag{34}$$

$$\left. \begin{array}{l} -a_i v_{i-1} + c_i v_i - b_i v_{i+1} = \varphi_i, \quad 1 \leqslant i \leqslant N-1, \\ v_0 = 1, \quad v_N = 0; \end{array} \right\} \tag{35}$$

$$\left. \begin{array}{l} -a_i w_{i-1} + c_i w_i - b_i w_{i+1} = \psi_i, \quad 1 \leqslant i \leqslant N-1, \\ w_0 = 0, \quad w_N = 1. \end{array} \right\} \tag{36}$$

Il s'ensuit de (33)-(36) que pour $1 \leqslant i \leqslant N-1$ les équations du système (32) ont une solution. Les conditions aux limites pour u_i, v_i et w_i garantissent la transformation de (33) en une identité pour $i = 0$ et $i = N$. Donc si les problèmes (34)-(36) sont résolus et y_0 et y_N trouvés, la formule (33) fournira la solution du problème initial (32). Cherchons d'abord y_0 et y_N.

On obtient les valeurs de y_0 et y_N en utilisant les équations du système (32) pour $i = 0$ et $i = N$. En portant dans ces équations y_i tiré de (33), on obtient le système de deux équations en y_0 et y_N:

$$\left(c_0 - \sum_{j=1}^{N-1} d_j v_j\right) y_0 - \left(\psi_0 + \sum_{j=1}^{N-1} d_j w_j\right) y_N = f_0 + \sum_{j=1}^{N-1} d_j u_j,$$

$$-\left(\varphi_N + \sum_{j=1}^{N-1} g_j v_j\right) y_0 + \left(c_N - \sum_{j=1}^{N-1} g_j w_j\right) y_N = f_N + \sum_{j=1}^{N-1} g_j u_j.$$

Si le déterminant de ce système

$$\Delta=\left(c_0-\sum_{j=1}^{N-1} d_j v_j\right)\left(c_N-\sum_{j=1}^{N-1} g_j w_j\right)-\left(\psi_0+\sum_{j=1}^{N-1} d_j w_j\right)\left(\varphi_N+\sum_{j=1}^{N-1} g_j v_j\right) \quad (37)$$

est différent de zéro, ce système possède une solution unique

$$y_0=\frac{1}{\Delta}\left[\left(c_N-\sum_{j=1}^{N-1} g_j w_j\right)\left(f_0+\sum_{j=1}^{N-1} d_j u_j\right)+\right.$$
$$\left.+\left(\psi_0+\sum_{j=1}^{N-1} d_j w_j\right)\left(f_N+\sum_{j=1}^{N-1} g_j u_j\right)\right], \quad (38)$$

$$y_N=\frac{1}{\Delta}\left[\left(\varphi_N+\sum_{j=1}^{N-1} g_j v_j\right)\left(f_0+\sum_{j=1}^{N-1} d_j u_j\right)+\right.$$
$$\left.+\left(c_0-\sum_{j=1}^{N-1} d_j v_j\right)\left(f_N+\sum_{j=1}^{N-1} g_j u_j\right)\right]. \quad (39)$$

Examinons maintenant la méthode de résolution des problèmes auxiliaires (34)-(36). Etant donné qu'on a affaire ici à des problèmes aux limites ordinaires pour équations triponctuelles, on est en mesure d'utiliser la méthode du balayage décrite au § 1. Pour (34)-(36) les formules de l'algorithme du balayage à droite prennent la forme suivante:

$$\begin{aligned} u_i &= \alpha_{i+1}u_{i+1}+\beta_{i+1}, \quad i=N-1,\ldots,0, \quad u_N=0,\\ v_i &= \alpha_{i+1}v_{i+1}+\gamma_{i+1}, \quad i=N-1,\ldots,0, \quad v_N=0,\\ w_i &= \alpha_{i+1}w_{i+1}+\delta_{i+1}, \quad i=N-1,\ldots,0, \quad w_N=1, \end{aligned} \quad (40)$$

où les coefficients de balayage α_i, β_i, γ_i et δ_i s'obtiennent sur la base des formules

$$\begin{aligned} &\alpha_{i+1}=\frac{b_i}{c_i-a_i\alpha_i}, \quad \beta_{i+1}=\frac{f_i+a_i\beta_i}{c_i-a_i\alpha_i},\\ &i=1, 2, \ldots, N-1, \quad \alpha_1=0, \quad \beta_1=0,\\ &\gamma_{i+1}=\frac{\varphi_i+a_i\gamma_i}{c_i-a_i\alpha_i}, \quad \delta_{i+1}=\frac{\psi_i+a_i\delta_i}{c_i-a_i\alpha_i},\\ &i=1, 2, \ldots, N-1, \quad \gamma_1=1, \quad \delta_1=0. \end{aligned} \quad (41)$$

Donc pour le problème (32) la méthode du balayage est décrite par les formules (33), (37)-(41).

Examinons maintenant la question de stabilité et de correction de l'algorithme proposé. En vertu du lemme 1 les conditions

$$|a_i|>0, \quad |b_i|>0, \quad |c_i|\geqslant|a_i|+|b_i|, \quad 1\leqslant i\leqslant N-1 \quad (42)$$

sont suffisantes pour garantir la stabilité et la correction de la méthode du balayage (40)-(41) mise en œuvre pour la résolution des problèmes auxiliaires (34)-(36). On peut montrer que si le système initial (32) possède une solution unique, le déterminant Δ, défini par la formule (37), est alors différent de zéro. Dans ce cas les formules (38) et (39), utilisées pour le calcul de y_0 et y_N, sont correctes. Formulons le résultat obtenu sous forme de lemme.

L e m m e 3. *Si le système* (32) *possède une solution unique et les conditions* (42) *sont remplies, l'algorithme* (33), (37)-(41) *de la méthode du balayage appliquée au problème* (32) *est correct et stable.*

Remarquons que la formulation de conditions suffisantes simples et en même temps pas trop limitatives de résolution du système (32) constitue un problème assez compliqué. Donnons un exemple des conditions garantissant à l'algorithme proposé des caractéristiques correctes et stables. Supposons que la matrice du système (32) est susceptible de dominance diagonale, c'est-à-dire que les conditions

$$|c_i| \geqslant |a_i| + |b_i| + |\varphi_i| + |\psi_i|, \quad 1 \leqslant i \leqslant N-1, \tag{43}$$

$$|c_0| \geqslant |\psi_0| + \sum_{j=1}^{N-1} |d_j|, \quad |c_N| \geqslant |\varphi_N| + \sum_{j=1}^{N-1} |g_j|, \tag{44}$$

$$|a_i| > 0, \quad |b_i| > 0, \quad 1 \leqslant i \leqslant N-1, \quad |c_0| > 0, \quad |c_N| > 0,$$

sont remplies de manière qu'au moins dans l'une des inégalités (43) ou (44) on ait une inégalité stricte.

Indiquons les principales étapes de la démonstration. On démontre d'abord qu'on a les inégalités $|\alpha_i| + |\gamma_i| + |\delta_i| \leqslant 1$, $1 \leqslant i \leqslant N$. Ensuite, on démontre les inégalités $|v_i| + |w_i| \leqslant 1$ pour $1 \leqslant i \leqslant N$, or, si dans (43) pour au moins un seul i l'inégalité stricte est vérifiée, alors pour tous les $1 \leqslant i \leqslant N$ est satisfaite l'inégalité $|v_i| + |w_i| < 1$. On a ensuite

$$\left|c_0 - \sum_{j=1}^{N-1} d_j v_j\right| \geqslant |c_0| - \sum_{j=1}^{N-1} |d_j|\,|v_j| \geqslant |\psi_0| +$$
$$+ \sum_{j=1}^{N-1} (1 - |v_j|)\,|d_j| \geqslant |\psi_0| + \sum_{j=1}^{N-1} |w_j|\,|d_j| \geqslant \left|\psi_0 + \sum_{j=1}^{N-1} w_j d_j\right|$$

et de façon analogue

$$\left|c_N - \sum_{j=1}^{N-1} g_j w_j\right| \geqslant \left|\varphi_N + \sum_{j=1}^{N-1} g_j v_j\right|,$$

et, au moins dans l'une de ces inégalités, on aboutit à une inégalité stricte. Il s'ensuit que le déterminant Δ défini dans (37) n'est pas nul. La stabilité et la correction de la méthode du balayage utilisée pour la résolution des problèmes auxiliaires (34)-(36) découlent de (43).

En guise d'exemple de problème se ramenant à (32), examinons le schéma pondéré

$$\begin{gathered} y_{t,\,i} = \sigma y^{n+1}_{\bar{x}x,\,i} + (1-\sigma)\, y^{n}_{\bar{x}x,\,i}, \quad 1 \leqslant i \leqslant N-1, \\ y^n_0 - y^n_k = \mu_1(t_n), \quad y^n_N - y^n_k = \mu_2(t_n), \\ y^0_i = u_0(x_i), \quad n = 0, 1, \ldots, \quad 1 \leqslant k \leqslant N-1, \end{gathered} \tag{45}$$

approximant l'équation de la conductibilité thermique aux conditions aux limites liées (non locales)

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < l, \quad t > 0,$$

$$u(0, t) - u(\nu(t), t) = \mu_1(t),$$

$$u(l, t) - u(\nu(t), t) = \mu_2(t), \quad u(x, 0) = u_0(x),$$

où la fonction $x = \nu(t)$ prend les valeurs de 0 à l. Notons que dans le schéma (45) la courbe $x = \nu(t)$ est approximée par la ligne brisée $x_k = \nu(t_n)$, de sorte que les points (x_k, t_n) constituent des points nodaux du réseau.

Le schéma aux différences (45) s'écrit sous forme de système (32) relativement à $y_i = y_i^{n+1}$ pour les valeurs suivantes des coefficients et du second membre $(\sigma \neq 0)$:

$$c_0=1, \quad d_k=1, \quad f_0=\mu_1(t_{n+1}), \quad \psi_0=0, \quad d_j=0, \quad j \neq k,$$

$$c_N=1, \quad q_k=1, \quad f_N=\mu_2(t_{n+1}), \quad \varphi_N=0, \quad g_j=0, \quad j \neq k,$$

$$\varphi_i=\psi_i=0, \quad a_i=b_i=1/h^2, \quad c_i=a_i+b_i+1/(\sigma\tau),$$

$$f_i=\frac{1}{\sigma\tau}\, y_i^n+\left(\frac{1}{\sigma}-1\right) y^n_{\bar{x}x,\, i}, \quad i=1, 2, \ldots, N-1.$$

De ces conditions on tire que l'exigence $|\, 2/h^2 + 1/(\sigma\tau)\,| > 2/h^2$ garantit la satisfaction des conditions (43), (44). Donc, avec $\sigma > -h^2/(4\tau)$, pour obtenir la solution du schéma (45) sur la couche supérieure, on peut appliquer la variante de la méthode du balayage décrite ici et qui dans ce cas sera stable et correcte.

4. Méthode de balayage non monotone. Revenons de nouveau à la résolution par la méthode du balayage des équations triponctuelles:

$$\begin{aligned} &c_0 y_0 - b_0 y_1 = f_0, && i = 0, \\ &-a_i y_{i-1} + c_i y_i - b_i y_{i+1} = f_i, && i = 1, 2, \ldots, N-1, \\ &-a_N y_{N-1} + c_N y_N = f_N, && i = N, \end{aligned} \qquad (46)$$

qu'on a élaborée au § 1. Rappelons que dans l'algorithme du balayage à droite (à gauche) les inconnues y_i s'obtiennent de proche en proche en allant dans le sens du décroissement (accroissement) de l'indice i. Dans ce cas y_i est exprimé au moyen de l'inconnue voisine. Cette structure de l'algorithme sert de fondement à l'appellation de la méthode construite de *méthode du balayage monotone*.

L'ordre monotone de détermination des inconnues y_i par remontée s'ensuit naturellement de l'ordre d'exclusion des inconnues des équations en sens direct. Donc la méthode du balayage monotone se réduit à la méthode d'exclusion de Gauss sans choix de l'élément principal et avec application à un système spécial d'équations algébriques linéaires (46) possédant une matrice tridiagonale. On sait qu'une telle variante de la méthode d'exclusion de Gauss est correcte au cas où le système d'équations possède une matrice à dominance diagonale. En ce qui concerne le système (46) cette assertion est démontrée au lemme 1.

Arrêtons-nous sur la question d'une manière plus détaillée. Rappelons qu'au § 1, point 1, à la l-ème opération d'exclusion des inconnues on avait obtenu dans le système (46) un système « rac-

courci »

$$
\begin{aligned}
&(c_l - a_l\alpha_l)\, y_l - b_l y_{l+1} = f_l + a_l \beta_l, && i = l, \\
&-a_i y_{i-1} + c_i y_i - b_i y_{i+1} = f_i, && l + 1 \leqslant i \leqslant N - 1, \qquad (47) \\
&-a_N y_{N-1} + c_N y_N = f_N
\end{aligned}
$$

d'inconnues $y_l, y_{l+1}, \ldots, y_N$. En posant $c_l - a_l\alpha_l$ différent de zéro, on a transformé la première équation du système (47) en la forme

$$y_l = \alpha_{l+1} y_{l+1} + \beta_{l+1}, \quad \alpha_{l+1} = b_l/(c_l - a_l\alpha_l) \tag{48}$$

et on l'a utilisée pour exclure y_l de l'équation (47) pour $i = l + 1$. Selon le lemme 1, si la matrice $\mathcal{A}$ du système (46) a une dominance diagonale, l'inégalité $|c_l - a_l\alpha_l| \geqslant |b_l|$ se vérifie. Donc dans la première équation du système (47) le coefficient associé à y_l est en module supérieur au coefficient associé à y_{l+1}. Aussi est-il inutile de choisir l'élément principal suivant la ligne, le passage à la forme (48) étant correct et la condition de stabilité $|\alpha_{l+1}| \leqslant 1$ se réalise automatiquement.

Si, par contre, la dominance diagonale n'a pas lieu, il devient impossible de garantir la non-nullité de la quantité $c_l - a_l\alpha_l$, de même que la vérité de l'inégalité $|\alpha_{l+1}| \leqslant 1$. Dans ce cas l'algorithme du balayage monotone peut engendrer une division par zéro ou une forte sensibilité aux erreurs d'arrondi, d'où nécessité de modifier un tel algorithme. La construction d'un algorithme correct de la méthode du balayage pour le système (46) possédant une solution unique s'appuie sur le choix de l'élément principal des lignes dans la méthode d'élimination de Gauss. Avec un tel algorithme l'ordre monotone de détermination des inconnues y_i peut être perturbé et c'est pourquoi cette méthode est appelée *méthode du balayage non monotone*.

Passons à la description de l'algorithme du balayage non monotone. Supposons qu'au bout de la l-ième opération d'exclusion de Gauss avec choix de l'élément principal des lignes, appliquée au système (46), on ait abouti au système « raccourci » suivant:

$$
\begin{aligned}
&Cy_{m_l} = b_l y_{l+1} = F, && i = l, \qquad (49) \\
&-Ay_{m_l} + c_{l+1} y_{l+1} - b_{l+1} y_{l+2} = \Phi, && i = l + 1, \qquad (50) \\
&-a_{l+2} y_{l+1} + c_{l+2} y_{l+2} - b_{l+2} y_{l+3} = f_{l+2}, && i = l + 2, \qquad (51) \\
&-a_i y_{i-1} + c_i y_i - b_i y_{i+1} = f_i, && l + 3 \leqslant i \leqslant N - 1, \qquad (52) \\
&-a_N y_{N-1} + c_N y_N = f_N, && i = N, \qquad (53)
\end{aligned}
$$

où $m_l \leqslant l$. (Pour $l = 0$ dans (49)-(53) il faut poser $C = c_0$, $A = a_1$, $F = f_0$, $\Phi = f_1$ et $m_0 = 0$.)

Décrivons la $(l+1)$-ième opération d'exclusion. La stratégie du choix de l'élément principal de la ligne se résout en deux cas:

a) Si $|C| \geqslant |b_l|$, l'équation (49) se transforme en

$$y_{m_l} - \alpha_{l+1} y_{l+1} = \beta_{l+1}, \quad \alpha_{l+1} = b_l/C, \quad \beta_{l+1} = F/C,$$

avec $|\alpha_{l+1}| \leqslant 1$, l'inconnue à indice m_l s'obtenant par l'intermédiaire de l'inconnue à indice $l+1$. Ensuite, avec l'équation obtenue on élimine y_{m_l} de (50). L'opération fournit l'équation suivante:

$$Cy_{m_{l+1}} - b_{l+1} y_{l+2} = F, \quad i = l+1, \tag{54}$$

où sont posés $m_{l+1} = l+1$, $C = c_{l+1} - A\alpha_{l+1}$, $F = \Phi + A\beta_{l+1}$. L'équation (51) n'est pas transformée, vu qu'elle contient y_{m_l}, mais est récrite sous forme

$$-Ay_{m_{l+1}} + C_{l+2} y_{l+2} - b_{l+2} y_{l+3} = \Phi, \quad i = l+2, \tag{55}$$

où l'on admet que $A = a_{l+2}$, $\Phi = f_{l+2}$. En joignant (54) et (55) à (52), (53), on obtient un nouveau système « raccourci » de la forme (49)-(53), où à l est substitué $l+1$. Sur ce point s'achève la $(l+1)$-ième opération.

b) Si $|C| < |b_l|$, (49) se transforme en

$$y_{l+1} - \alpha_{l+1} y_{m_l} = \beta_{l+1}, \quad \alpha_{l+1} = C/b_l, \quad \beta_{l+1} = -F/b_l,$$

où de nouveau $|\alpha_{l+1}| \leqslant 1$, mais cette fois l'inconnue à indice $l+1$ se calcule par l'intermédiaire de l'inconnue à indice m_l. L'équation obtenue est utilisée à des fins d'exclusion de y_{l+1} de (50) et (51). Dans ce cas l'équation (50) est transformée en la forme (54), où $m_{l+1} = m_l$, $C = c_{l+1}\alpha_{l+1} - A$, $F = \Phi - c_{l+1}\beta_{l+1}$, et l'équation (51) en la forme (55), où A et Φ sont redéfinis selon les formules $A = a_{l+2}\alpha_{l+1}$, $\Phi = f_{l+2} + a_{l+2}\beta_{l+1}$. Les équations (52), (53) ne sont pas transformées, car elles ne contiennent pas y_{l+1}. On obtient de nouveau le système de la forme (49)-(53). Il diffère de celui obtenu dans le premier cas par les coefficients C et A et les seconds membres F et Φ calculés au moyen d'autres formules.

Bref, on a décrit une opération d'élimination avec choix de l'élément principal. Remarquons que si le système primitif n'est pas dégénéré, dans l'équation (49) les coefficients C et b_l ne peuvent devenir nuls simultanément. Cela garantit la correction des formules permettant d'obtenir les coefficients de balayage α_{l+1} et β_{l+1}. Vu que tous les α_{l+1} calculés sont en module inférieurs à l'unité, le calcul des inconnues y_i par remontée s'avère stable relativement aux erreurs d'arrondi.

Avec l'algorithme proposé l'ordre de calcul des inconnues peut être de nature non monotone. On est alors obligé de mémoriser l'information sur l'inconnue calculée par l'intermédiaire de l'inconnue déjà trouvée à l'aide des coefficients α_{i+1} et β_{i+1} au cours des opérations précédentes. Cette information peut être stockée sous forme de deux ensembles d'indices entiers θ et $\varkappa: \theta = \{\theta_i, 1 \leqslant i \leqslant N\}$, $\varkappa = \{\varkappa_i, 1 \leqslant i \leqslant N\}$, de sorte que les inconnues s'obtiennent par les formules $y_m = \alpha_{i+1}y_n + \beta_{i+1}$, $m = \theta_{i+1}$, $n = \varkappa_{i+1}$, $i = N - 1, N - 2, \ldots, 0$. Les ensembles θ et $\varkappa$ sont construits avec la méthode utilisée en sens direct.

L'algorithme de la méthode du balayage non monotone peut être déterminé complètement de la façon suivante.

1) On fixe les valeurs initiales de C, A, F et Φ: $C = c_0$, $A = a_1$, $F = f_0$, $\Phi = f_1$ et l'on pose de façon formelle $\varkappa_0 = 0$.

2) On effectue successivement pour $i = 0, 1, \ldots, N - 1$, selon la situation, les opérations décrites aux points a) ou b):

a) si $|C| \geqslant |b_i|$, $\alpha_{i+1} = b_i/C$, $\beta_{i+1} = F/C$, $C = c_{i+1} - A\alpha_{i+1}$, $F = \Phi + A\beta_{i+1}$, $\theta_{i+1} = \varkappa_i$, $\varkappa_{i+1} = i + 1$, $A = a_{i+2}$, $\Phi = f_{i+2}$;

b) si $|C| < |b_i|$, $\alpha_{i+1} = C/b_i$, $\beta_{i+1} = -F/b_i$, $C = c_{i+1}\alpha_{i+1} - A$, $F = \Phi - c_{i+1}\beta_{i+1}$, $\theta_{i+1} = i + 1$, $\varkappa_{i+1} = \varkappa_i$, $A = a_{i+2}\alpha_{i+1}$, $\Phi = f_{i+2} + a_{i+2}\beta_{i+1}$.

R e m a r q u e. Pour $i = N - 1$, il ne faut plus effectuer la redéfinition de A et Φ aux points a) ou b).

3) On calcule d'abord l'inconnue y_n, où $n = \varkappa_N$ selon la formule $y_n = F/C$, ensuite, successivement pour $i = N - 1, N - 2, \ldots, 0$ on calcule les autres inconnues $y_m = \alpha_{i+1}y_n + \beta_{i+1}$, $m = \theta_{i+1}$, $n = \varkappa_{i+1}$.

Remarquons que l'algorithme proposé ici se transforme en un algorithme ordinaire du balayage à droite si les conditions du lemme 1 sont remplies.

Un calcul élémentaire du nombre d'opérations arithmétiques impliquant l'obtention de l'algorithme de la méthode du balayage non monotone montre qu'au pis aller, quand pour tout i les calculs sont exécutés selon les formules du point b), il faut $Q = 12N$ opérations. C'est 1,5 fois supérieur au nombre d'opérations exigé par l'algorithme du balayage monotone.

Voyons un exemple d'application de la méthode du balayage non monotone. Supposons qu'il s'agit de résoudre le problème de différences suivant:

$$\begin{gathered} -y_{i-1} + y_i - y_{i+1} = 0, \quad 1 \leqslant i \leqslant N - 1, \\ y_0 = 1, \quad y_N = 0. \end{gathered} \tag{56}$$

Le problème (56) est un cas particulier du système (46), où $f_0 = 1$, $b_0 = a_N = 0$, $c_0 = c_N = 1$, $f_N = 0$, $c_i = a_i = b_i = 1$, $f_i = 0$, $1 \leqslant i \leqslant N - 1$. Si N n'est pas multiple de 3, le problème (56)

admet une solution de la forme (voir point 1, § 4, ch. I)

$$y_i = \sin\frac{(N-i)\pi}{3} \Big/ \sin\frac{N\pi}{3}, \quad 0 \leqslant i \leqslant N. \tag{57}$$

Les algorithmes des balayages à droite et à gauche appliqués à (56) sont incorrects, car lors du calcul des coefficients de balayage, α_3 pour le balayage à droite et ξ_{N-2} pour le balayage à gauche, on est obligé d'effectuer la division par un dénominateur nul $c_2 - a_2\alpha_2$ ou $c_{N-2} - b_{N-2}\xi_{N-1}$. Par contre, l'algorithme du balayage non monotone permet d'obtenir une solution exacte pour (57). Donnons en guise d'illustration (tabl. 1) les valeurs des coefficients α_i, β_i, ainsi que de θ_i et $\varkappa_i$ pour $N = 11$.

Tableau 1

i	0	1	2	3	4	5	6	7	8	9	10	11
α_i		0	1	0	−1	1	0	−1	1	0	−1	1
β_i		1	1	−1	−1	−1	1	1	1	−1	−1	−1
θ_i		0	1	3	2	4	6	5	7	9	8	10
$\varkappa_i$		1	2	2	4	5	5	7	8	8	10	11
y_i	1	1	0	−1	−1	0	1	1	0	−1	−1	0

§ 3. Méthode du balayage pour les équations pentaponctuelles

1. Algorithme du balayage monotone. On a passé plus haut en revue les différentes variantes de la méthode du balayage utilisée à des fins de résolution des équations aux différences triponctuelles. Comme il a déjà été mentionné, ces équations aux différences surgissent lors de l'approximation des problèmes aux limites sur les équations différentielles de second ordre.

Pour la recherche de la solution de problèmes aux limites sur des équations d'un ordre plus élevé on peut choisir deux procédés. Le premier consiste à passer au système d'équations différentielles du premier ordre et à construire le schéma aux différences correspondant. On obtient dans ce cas le problème aux limites sur les équations vectorielles biponctuelles. La méthode de résolution de ces problèmes sera exposée au § 4.

Le second procédé consiste dans l'approximation directe du problème différentiel posé. Dans ce cas on aboutit à des équations aux différences multiponctuelles. On se heurte le plus souvent aux

systèmes d'équations pentaponctuelles de la forme suivante:

$$c_0y_0 - d_0y_1 + e_0y_2 = f_0, \qquad i = 0, \tag{1}$$

$$-b_1y_0 + c_1y_1 - d_1y_2 + e_1y_3 = f_1, \qquad i = 1, \tag{2}$$

$$a_iy_{i-2} - b_iy_{i-1} + c_iy_i - d_iy_{i+1} + e_iy_{i+2} = f_i, \quad 2 \leqslant i \leqslant N-2, \tag{3}$$

$$a_{N-1}y_{N-3} - b_{N-1}y_{N-2} + c_{N-1}y_{N-1} - d_{N-1}y_N = f_{N-1}, \qquad i = N-1, \tag{4}$$

$$a_Ny_{N-2} - b_Ny_{N-1} + c_Ny_N = f_N, \qquad i = N. \tag{5}$$

Ces formes de systèmes apparaissent lors de l'approximation des problèmes aux limites pour les équations différentielles du quatrième ordre, ainsi que lors de la construction de schémas aux différences pour des équations aux dérivées partielles. La matrice $\mathcal{A}$ du système (1)-(5) est une matrice carrée pentadiagonale de dimension $(N + 1) \times \times (N + 1)$ et possède au plus $5N - 1$ éléments non nuls.

Pour la résolution du système (1)-(5), recourrons à la méthode d'élimination de Gauss. Compte tenu de la structure du système (1)-(5), on constate sans peine que les formules d'élimination par remontée de la méthode de Gauss sont:

$$y_i = \alpha_{i+1}y_{i+1} - \beta_{i+1}y_{i+2} + \gamma_{i+1}, \quad 0 \leqslant i \leqslant N-2, \tag{6}$$

$$y_{N-1} = \alpha_Ny_N + \gamma_N, \qquad i = N-1. \tag{7}$$

La mise en œuvre de (6), (7) oblige à fixer y_N, ainsi qu'à déterminer les coefficients α_i, β_i, γ_i.

Cherchons d'abord les formules pour α_i, β_i, et γ_i. En utilisant (6), exprimons y_{i-1} et y_{i-2} au moyen de y_i et y_{i+1}. Il vient

$$y_{i-1} = \alpha_iy_i - \beta_iy_{i+1} + \gamma_i, \quad 1 \leqslant i \leqslant N-1, \tag{8}$$

$$y_{i-2} = (\alpha_i\alpha_{i-1} - \beta_{i-1})\, y_i - \beta_i\alpha_{i-1}y_{i+1} + \alpha_{i-1}\gamma_i + \gamma_{i-1} \tag{9}$$

pour $2 \leqslant i \leqslant N-1$.

En portant (8) et (9) dans (3), il vient

$$[c_i - a_i\beta_{i-1} + \alpha_i\,(a_i\alpha_{i-1} - b_i)]\, y_i = [d_i + \beta_i\,(a_i\alpha_{i-1} - b_i)]\, y_{i+1} - \\ - e_iy_{i+2} + [f_i - a_i\gamma_{i-1} - \gamma_i\,(a_i\alpha_{i-1} - b_i)], \quad 2 \leqslant i \leqslant N-2.$$

En comparant cette expression à (6), on voit que si l'on pose

$$\alpha_{i+1} = \frac{1}{\Delta_i}[d_i + \beta_i\,(a_i\alpha_{i-1} - b_i)], \quad \beta_{i+1} = \frac{e_i}{\Delta_i}, \\ \gamma_{i+1} = \frac{1}{\Delta_i}[f_i - a_i\gamma_{i-1} - \gamma_i\,(a_i\alpha_{i-1} - b_i)], \tag{10}$$

où $\Delta_i = c_i - a_i\beta_{i-1} + \alpha_i\,(a_i\alpha_{i-1} - b_i)$, les équations du système (1)-(5) seront vérifiées pour $2 \leqslant i \leqslant N-2$.

Les relations de récurrence (10) lient α_{i+1}, β_{i+1} et γ_{i+1} à α_i, α_{i-1}, β_i, β_{i-1}, γ_i et γ_{i-1}. Aussi, si α_i, β_i et γ_i sont donnés pour $i = 1, 2$, alors, à l'aide des formules (10), on peut trouver successivement les coefficients α_i, β_i et γ_i pour $3 \leqslant i \leqslant N-1$.

Cherchons α_i, β_i et γ_i pour $i = 1, 2$. De (1) et de (6) on obtient directement pour $i = 0$

$$\alpha_1 = d_0/c_0, \quad \beta_1 = e_0/c_0, \quad \gamma_1 = f_0/c_0. \tag{11}$$

Ensuite, en portant les valeurs de (8) pour $i = 1$ dans (2), l'on obtient

$$(c_1 - b_1\alpha_1)\, y_1 = (d_1 - b_1\beta_1)\, y_2 - e_1 y_3 + f_1 + b_1\gamma_1.$$

Donc (2) sera vérifié si l'on pose

$$\alpha_2 = \frac{d_1 - b_1\beta_1}{c_1 - b_1\alpha_1}, \quad \beta_2 = \frac{e_1}{c_1 - b_1\alpha_1}, \quad \gamma_2 = \frac{f_1 + b_1\gamma_1}{c_1 - b_1\alpha_1}. \tag{12}$$

Bref, en utilisant (10)-(12), on peut trouver α_i, β_i et γ_i pour $1 \leqslant i \leqslant N-1$. Il reste à déterminer α_N, γ_N et y_N entrant dans la formule (7).

Utilisons pour cela les équations (4) et (5). En portant (8) et (9) pour $i = N-1$ dans (4) et en comparant l'expression obtenue avec (7), on trouve que α_N et γ_N se déterminent selon les formules (10) pour $i = N-1$. Cherchons maintenant y_N. A cette fin portons (6) pour $i = N-2$ et (7) dans l'équation (5). Il vient

$$[c_N - a_N\beta_{N-1} + \alpha_N (a_N\alpha_{N-1} - b_N)]\, y_N = \\ = f_N - a_N\gamma_{N-1} - \gamma_N (a_N\alpha_{N-1} - b_N)$$

ou

$$y_N = \gamma_{N+1},$$

où γ_{N+1} se détermine selon la formule (10) pour $i = N$.

En réunissant les formules obtenues précédemment, écrivons l'algorithme du balayage à droite pour le système (1)-(5) sous la forme suivante:

1) selon les formules

$$\alpha_{i+1} = \frac{1}{\Delta_i}[d_i + \beta_i (a_i\alpha_{i-1} - b_i)], \quad i = 2, 3, \ldots, N-1, \tag{13}$$

$$\alpha_1 = \frac{d_0}{c_0}, \quad \alpha_2 = \frac{1}{\Delta_1}(d_1 - \beta_1 b_1),$$

$$\gamma_{i+1} = \frac{1}{\Delta_i}[f_i - a_i\gamma_{i-1} - \gamma_i (a_i\alpha_{i-1} - b_i)], \quad i = 2, 3, \ldots, N, \tag{14}$$

$$\gamma_1 = \frac{f_0}{c_0}, \quad \gamma_2 = \frac{1}{\Delta_1}(f_1 + b_1\gamma_1),$$

$$\beta_{i+1} = e_i/\Delta_i, \quad i = 1, 2, \ldots, N-2, \quad \beta_1 = e_0/c_0, \tag{15}$$

où

$$\Delta_i = c_i - a_i\beta_{i-1} + \alpha_i (a_i\alpha_{i-1} - b_i), \quad 2 \leqslant i \leqslant N,$$
$$\Delta_1 = c_1 - b_1\alpha_1, \quad (16)$$

on obtient les coefficients de balayage α_i, β_i et γ_i;

2) les inconnues y_i s'obtiennent de proche en proche selon les formules

$$y_i = \alpha_{i+1}y_{i+1} - \beta_{i+1}y_{i+2} + \gamma_{i+1}, \quad i = N-2, N-3, \ldots, 0,$$
$$y_{N-1} = \alpha_N y_N + \gamma_N, \quad y_N = \gamma_{N+1}. \quad (17)$$

L'algorithme construit sera également appelé *algorithme du balayage monotone*.

R e m a r q u e. On construira sans peine l'algorithme du balayage à gauche, de même que l'algorithme des balayages opposés pour le système (1)-(5).

Calculons le nombre d'opérations arithmétiques exigé par l'algorithme (13)-(17). Pour la mise en œuvre de (13)-(17) il faudra $8N - 5$ additions et soustractions, $8N - 5$ multiplications et $3N$ divisions. Abstraction faite du temps mis aux différentes opérations arithmétiques par l'ordinateur, le nombre total d'opérations associé à l'algorithme proposé vaut $Q = 19N - 10$.

2. Justification de la méthode. L'algorithme du balayage (13)-(17) élaboré plus haut sera dit *correct* si, pour tout $2 \leqslant i \leqslant N$, on aura l'inégalité

$$\Delta_i = c_i - a_i\beta_{i-1} + \alpha_i (a_i\alpha_{i-1} - b_i) \neq 0, \qquad \Delta_1 = c_1 - \alpha_1 b_1 \neq 0.$$

Le lemme suivant fournit les conditions suffisantes de la correction de l'algorithme (13)-(17).

L e m m e 4. *Soient les coefficients du système* (1)-(5) *satisfaisant aux conditions*

$$|a_i| > 0, \quad 2 \leqslant i \leqslant N, \qquad |b_i| > 0, \quad 1 \leqslant i \leqslant N,$$
$$|d_i| > 0, \quad 0 \leqslant i \leqslant N-1, \quad |e_i| > 0, \quad 0 \leqslant i \leqslant N-2,$$

ainsi qu'aux conditions

$$|c_0| \geqslant |d_0| + |e_0|, \quad |c_1| \geqslant |b_1| + |d_1| + |e_1|,$$
$$|c_N| \geqslant |a_N| + |b_N|, \quad |c_{N-1}| \geqslant |a_{N-1}| + |b_{N-1}| + |d_{N-1}|, \quad (18)$$
$$|c_i| \geqslant |a_i| + |b_i| + |d_i| + |e_i|, \quad 2 \leqslant i \leqslant N-2,$$

avec au moins une inégalité stricte dans l'une des inégalités (18). *Dans ce cas l'algorithme* (13)-(17) *est correct et, en outre, on a les inégalités*

$$|\alpha_i| + |\beta_i| \leqslant 1, \quad 1 \leqslant i \leqslant N-1, \quad |\alpha_N| \leqslant 1.$$

En effet, en vertu des conditions du lemme, il s'ensuit de (13) et (15)

$$|\alpha_1| + |\beta_1| = \frac{|d_0| + |e_0|}{|c_0|} \leqslant 1.$$

Ensuite, utilisant l'inégalité obtenue $1 - |\alpha_1| \geqslant |\beta_1|$, on trouve

$$|c_1 - b_1\alpha_1| \geqslant |c_1| - |b_1| |\alpha_1| \geqslant |b_1| (1 - |\alpha_1|) + \\ + |d_1| + |e_1| \geqslant |b_1| |\beta_1| + |d_1| + \\ + |e_1| \geqslant |d_1 - b_1\beta_1| + |e_1| > 0.$$

De cette inégalité et de (13)-(15) on obtient la relation

$$|\alpha_2| + |\beta_2| = \frac{|d_1 - \beta_1 b_1| + |e_1|}{|c_1 - b_1\alpha_1|} \leqslant 1.$$

La suite de la démonstration se fera par induction. Soient satisfaites les inégalités

$$|\alpha_{i-1}| + |\beta_{i-1}| \leqslant 1, \quad |\alpha_i| + |\beta_i| \leqslant 1. \tag{19}$$

Montrons qu'alors les inégalités

$$\Delta_i = c_i - a_i\beta_{i-1} + \alpha_i (a_i\alpha_{i-1} - b_i) \neq 0, \quad |\alpha_{i+1}| + |\beta_{i+1}| \leqslant 1$$

sont vérifiées.

Et, de fait, de (18) et de (19) il vient

$$|\Delta_i| \geqslant |c_i| - |a_i| |\beta_{i-1}| - |\alpha_i| |\alpha_{i-1}| |a_i| - \\ - |\alpha_i| |b_i| \geqslant |a_i| (1 - |\beta_{i-1}|) + |b_i| (1 - |\alpha_i|) - \\ - |\alpha_i| |\alpha_{i-1}| |a_i| + |d_i| + |e_i| \geqslant |a_i| |\alpha_{i-1}| + \\ + |b_i| |\beta_i| - |\alpha_i| |\alpha_{i-1}| |a_i| + |d_i| + |e_i| \geqslant \\ \geqslant |a_i| |\alpha_{i-1}| (1 - |\alpha_i|) + |d_i - b_i\beta_i| + |e_i| \geqslant \\ \geqslant |a_i| |\alpha_{i-1}| |\beta_i| + |d_i - b_i\beta_i| + |e_i| \geqslant \\ \geqslant |d_i + \beta_i (a_i\alpha_{i-1} - b_i)| + |e_i| > 0, \quad i \leqslant N-2. \tag{20}$$

De là et de (13), (15) on tire

$$|\alpha_{i+1}| + |\beta_{i+1}| = \frac{|d_i + \beta_i (a_i\alpha_{i-1} - b_i)| + |e_i|}{|\Delta_i|} \leqslant 1, \quad i \leqslant N-2.$$

Ensuite, pour $i = N - 1$, on a au lieu de (20) l'estimation

$$|\Delta_{N-1}| \geqslant |a_{N-1}|\,|\alpha_{N-2}|\,|\beta_{N-1}| + |b_{N-1}|\,|\beta_{N-1}| + + |d_{N-1}| > 0.$$

De plus, on obtient de là

$$|\Delta_{N-1}| \geqslant |d_{N-1} + \beta_{N-1}(a_{N-1}\alpha_{N-2} - b_{N-1})|,$$

et, par suite,

$$|\alpha_N| = \frac{1}{|\Delta_{N-1}|}\,|d_{N-1} + \beta_{N-1}(a_{N-1}\alpha_{N-2} - b_{N-1})| \leqslant 1.$$

Il ne reste qu'à montrer que $\Delta_N \neq 0$. On aura

$$|\Delta_N| \geqslant |c_N| - |a_N|\,|\beta_{N-1}| - |\alpha_N|\,|\alpha_{N-1}|\,|a_N| - - |\alpha_N|\,|b_N| = |c_N| - |a_N| - |b_N| + |a_N|(1 - - |\beta_{N-1}|) + |b_N|(1 - |\alpha_N|) - |\alpha_N|\,|\alpha_{N-1}|\,|a_N| \geqslant \geqslant |c_N| - |a_N| - |b_N| + (1 - |\alpha_N|)(1 - |\beta_{N-1}|)\,|a_N| + + |b_N|(1 - |\alpha_N|).$$

En vertu des hypothèses du lemme, on obtient sans peine qu'au moins dans l'une des inégalités $|c_N| \geqslant |a_N| + |b_N|$, $|\alpha_N| \leqslant 1$, on aboutit à une inégalité stricte. Il s'ensuit donc que $\Delta_N \neq 0$. Le lemme est démontré.

R e m a r q u e. Il s'ensuit des estimations $|\alpha_i| + |\beta_i| \leqslant 1$ données dans le lemme 4 que si, lors du calcul de y_N une erreur est commise, cette dernière ne croîtra pas en poursuivant le calcul suivant les formules (17).

3. Variante du balayage non monotone. Donnons maintenant l'algorithme de la méthode du balayage. obtenu si l'on recherche la solution du système (1)-(5) suivant la méthode de Gauss avec choix de l'élément principal de la ligne. Cet algorithme sera correct à la seule condition de non-dégénérescence de la matrice $\mathcal{A}$ du système (1)-(5). Vu que le procédé de construction de l'algorithme est analogue à celui décrit au point 4 du § 2, on se limitera à ne donner que la forme définitive de l'algorithme.

1) On fournit les valeurs initiales: $C = c_0$, $D = d_0$, $B = b_1$, $Q = c_1$, $S = a_2$, $T = b_2$, $R = 0$, $A = a_3$, $F = f_0$, $\Phi = f_1$, $G = f_2$, $H = f_3$ en posant $\varkappa_0 = 0$, $\eta_0 = 1$.

2) On effectue successivement pour $i = 0, 1, \ldots, N - 2$, suivant la situation, les opérations décrites aux points a), b) ou c):

a) si $|C| \geqslant |D|$ et $|C| \geqslant |e_i|$, on a alors

$$\begin{gathered}\alpha_{i+1} = D/C, \quad \beta_{i+1} = e_i/C, \quad \gamma_{i+1} = F/C,\\ C = Q - B\alpha_{i+1}, \quad D = d_{i+1} - B\beta_{i+1}, \quad F = \Phi + B\gamma_{i+1},\\ B = T - S\alpha_{i+1}, \quad Q = c_{i+2} - S\beta_{i+1}, \quad \Phi = G - S\gamma_{i+1},\\ S = A - R\alpha_{i+1}, \quad T = b_{i+3} - R\beta_{i+1}, \quad G = H + R\gamma_{i+1},\end{gathered} \qquad (21)$$

$$\left.\begin{gathered}R = 0, \quad A = a_{i+4}, \quad H = f_{i+4},\\ \theta_{i+1} = \varkappa_i, \quad \varkappa_{i+1} = \eta_i, \quad \eta_{i+1} = i + 2;\end{gathered}\right\} \qquad (22)$$

b) si $|D| > |C|$ et $|D| \geqslant |e_i|$, on a alors

$$\alpha_{i+1} = C/D, \quad \beta_{i+1} = -e_i/D, \quad \gamma_{i+1} = -F/D,$$
$$C = Q\alpha_{i+1} - B, \quad D = Q\beta_{i+1} + d_{i+1}, \quad F = \Phi - Q\gamma_{i+1},$$
$$B = T\alpha_{i+1} - S, \quad Q = T\beta_{i+1} + c_{i+2}, \quad \Phi = T\gamma_{i+1} + G,$$
$$S = A\alpha_{i+1} - R, \quad T = A\beta_{i+1} + b_{i+3}, \quad G = H - A\gamma_{i+1}, \qquad (23)$$
$$\left.\begin{array}{c} R=0, \quad A=a_{i+4}, \quad H=f_{i+4}, \\ \theta_{i+1}=\eta_i, \quad \varkappa_{i+1}=\varkappa_i, \quad \eta_{i+1}=i+2; \end{array}\right\} \qquad (24)$$

c) si $|e_i| > |C|$ et $|e_i| > |D|$, on a alors

$$\alpha_{i+1} = D/e_i, \quad \beta_{i+1} = C/e_i, \quad \gamma_{i+1} = F/e_i,$$
$$C = Q - d_{i+1}\alpha_{i+1}, \quad D = B - d_{i+1}\beta_{i+1}, \quad F = \Phi + d_{i+1}\gamma_{i+1},$$
$$B = T - c_{i+2}\alpha_{i+1}, \quad Q = S - c_{i+2}\beta_{i+1}, \quad \Phi = G - c_{i+2}\gamma_{i+1},$$
$$S = A - b_{i+3}\alpha_{i+1}, \quad T = R - b_{i+3}\beta_{i+1}, \quad G = H + b_{i+3}\gamma_{i+1}, \qquad (25)$$
$$\left.\begin{array}{c} R = -a_{i+4}\alpha_{i+1}, \quad A = -a_{i+4}\beta_{i+1}, \quad H = f_{i+4} - a_{i+4}\gamma_{i+1}, \\ \theta_{i+1} = i+2, \quad \varkappa_{i+1} = \eta_i, \quad \eta_{i+1} = \varkappa_i. \end{array}\right\} \qquad (26)$$

Remarque. Pour $i \geqslant N-3$ on peut se dispenser d'effec·uer les calculs suivant les formules (22), (24) ou (26), tandis que pour $i = N-2$ on n'effectue pas les calculs suivant les formules (21), (23), (25).

3) Si $|C| \geqslant |D|$, on a $\alpha_N = D/C$, $\gamma_N = F/C$, $\gamma_{N+1} = (\Phi + B\gamma_N)/(Q - B\alpha_N)$, $\theta_N = \varkappa_{N-1}$, $\varkappa_N = \eta_{N-1}$. Si $|D| > |C|$, on a $\alpha_N = C/D$, $\gamma_N = -F/D$, $\gamma_{N+1} = (\Phi - Q\gamma_N)/(Q\alpha_N - B)$, $\theta_N = \eta_{N-1}$, $\varkappa_N = \varkappa_{N-1}$.

4) On calcule les inconnues $y_n = \gamma_{N+1}$, $y_m = \alpha_N y_n + \gamma_N$, $m = \theta_N$, $n = \varkappa_N$, ensuite, on détermine successivement pour $i = N-2, N-3, \ldots, 0$ les inconnues restantes $y_m = \alpha_{i+1}y_n - \beta_{i+1}y_k + \gamma_{i+1}$, $m = \theta_{i+1}$, $n = \varkappa_{i+1}$, $k = \eta_{i+1}$.

Voyons un exemple d'application de la méthode du balayage non monotone. Au point 3 du § 3, ch. I on a résolu le problème aux limites discret suivant:

$$\begin{array}{ll} y_0 - y_1 + 2y_2 = 2, & i = 0, \\ -y_0 + y_1 - y_2 + y_3 = 0, & i = 1, \\ y_{i-2} - y_{i-1} + 2y_i - y_{i+1} + y_{i+2} = 0, & 2 \leqslant i \leqslant N-2, \\ y_{N-3} - y_{N-2} + y_{N-1} - y_N = 0, & i = N-1, \\ 2y_{N-2} - y_{N-1} + y_N = 0, & i = N. \end{array} \qquad (27)$$

Si N est pair et n'est pas multiple de 3, le système (27) possède une solution unique

$$y_i = -\cos\frac{i\pi}{2} - \sin\frac{i\pi}{2}, \quad 0 \leqslant i \leqslant N. \qquad (28)$$

On se convainc sans peine que l'algorithme du balayage monotone construit pour le système (27) est incorrect: lors du calcul des coefficients de balayage α_2, β_2 et γ_2 on est obligé d'effectuer une division par zéro. Par contre, l'algorithme du balayage non monotone permet d'obtenir une solution exacte pour (28). En guise d'illustration de cet algorithme (tabl. 2), donnons les valeurs des coefficients de balayage α_i, β_i et γ_i, de même que de θ_i, $\varkappa_i$ et η_i pour $N = 10$.

Tableau 2

i	0	1	2	3	4	5	6	7	8	9	10	11
α_i		$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	0	0	0	$-\frac{1}{3}$	$-\frac{1}{3}$	0	1	
β_i		$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	1	-1	1	$-\frac{2}{3}$	$-\frac{2}{3}$	1		
γ_i		1	1	-1	-2	-2	2	$\frac{4}{3}$	$-\frac{2}{3}$	0	-2	1
θ_i		2	3	4	0	5	6	7	9	8	1	
$\varkappa_i$		1	0	1	1	1	1	1	8	1	10	
η_i		0	1	0	5	6	7	8	1	10		
y_i	-1	-1	1	1	-1	-1	1	1	-1	-1	1	

On voit sur le tableau que les inconnues y_i se déterminent dans l'ordre suivant: y_{10}, y_1, y_8, y_9, y_7, y_6, y_5, y_0, y_4, y_3, y_2, c'est-à-dire suivant un ordre non monotone.

§ 4. Méthode du balayage matriciel

1. Systèmes d'équations vectorielles. On a précédemment noté que l'un des procédés de résolution des problèmes aux limites sur les équations différentielles ordinaires d'ordre élevé est leur réduction à un système d'équations du premier ordre avec approximation subséquente de ce système par un schéma aux différences. On obtient ainsi un *système d'équations vectoriel biponctuel* de conditions aux limites de première espèce. Sous forme générale il s'écrit ainsi:

$$\begin{gathered} P_{i+1}V_{i+1} - Q_iV_i = F_{i+1}, \quad 0 \leqslant i \leqslant N-1, \\ P_0V_0 = F_0, \quad Q_NV_N = F_{N+1}, \end{gathered} \tag{1}$$

où V_i est le vecteur des inconnues de dimension M, P_{i+1} et Q_i, pour $0 \leqslant i \leqslant N-1$, des matrices carrées $M \times M$, P_0 et Q_N des matrices rectangulaires de dimensions $M_1 \times M$ et $M_2 \times M$ respectivement, $M_1 + M_2 = M$. Le vecteur F_{i+1} pour $0 \leqslant i \leqslant N-1$ est de dimension M, tandis que les vecteurs F_0 et F_{N+1} le sont de M_1 et M_2 respectivement.

Notons que l'autre procédé de résolution des équations différentielles mentionnées est leur approximation directe par des schémas aux différences. On obtiendra alors un système d'équations scalaires multiponctuelles. On a déjà étudié les méthodes de résolution des équations scalaires tri- et pentaponctuelles au § 1-3. Mais

si l'on procède à l'approximation d'un système d'équations différentielles ordinaires d'ordre élevé, on aboutit alors à un système d'équations vectorielles multiponctuelles. Cependant, les systèmes d'équations multiponctuelles vectorielles, comme les systèmes d'équations multiponctuelles scalaires, peuvent être réduits aux systèmes de la forme (1). Et à chaque méthode de résolution de (1) correspondra une certaine méthode de résolution du système multiponctuel initial. Eclairons l'idée de la transformation mentionnée sur l'exemple d'un système d'équations pentaponctuelles déjà étudié au § 3 (voir (1)-(5)). Si l'on pose

$$\begin{aligned} Y_i &= (y_{i+1},\ y_i,\ y_{i-1},\ y_{i-2}), && 2 \leqslant i \leqslant N-1, \\ F_{i+1} &= (f_i,\ 0,\ 0,\ 0), && 2 \leqslant i \leqslant N-2, \\ F_2 &= (f_0,\ f_1), \quad F_N = (f_{N-1},\ f_N), \end{aligned}$$

alors compte tenu des relations identiques entre Y_{i+1} et Y_i le système considéré du § 3 s'écrira sous la forme

$$\begin{aligned} &P_{i+1}Y_{i+1} - Q_iY_i = F_{i+1}, \quad 2 \leqslant i \leqslant N-2, \\ &P_2Y_2 = F_2, \quad Q_{N-1}Y_{N-1} = F_N, \end{aligned} \tag{2}$$

où

$$P_{i+1} = \begin{Vmatrix} e_i & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{Vmatrix}, \quad Q_i = \begin{Vmatrix} d_i & -c_i & b_i & -a_i \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{Vmatrix},$$

$$2 \leqslant i \leqslant N-2,$$

$$P_2 = \begin{Vmatrix} 0 & e_0 & -d_0 & c_0 \\ e_1 & -d_1 & c_1 & -b_1 \end{Vmatrix}, \quad Q_{N-1} = \begin{Vmatrix} -d_{N-1} & c_{N-1} & -b_{N-1} & a_{N-1} \\ c_N & -b_N & a_N & 0 \end{Vmatrix}.$$

Dans le cas considéré $M_1 = M_2 = 2$, $M = 4$.

Nonobstant le fait que les équations vectorielles multiponctuelles peuvent être ramenées à la forme (1) et qu'on peut ainsi se limiter à la construction de la méthode de résolution du seul système (1), on examinera séparément la classe des *équations vectorielles triponctuelles*. Et, même davantage, au point 3 on réduira (1) à un système d'équations vectorielles triponctuelles et l'on obtiendra une méthode de résolution du système (1) constituant une variante de la méthode de résolution d'équations triponctuelles.

Avant de décrire la forme générale des équations triponctuelles, donnons un exemple. On montrera que le problème de différences pour la plus simple des équations elliptiques se réduit à un système d'équations triponctuelles de forme spéciale.

Soit un maillage rectangulaire $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leqslant \leqslant i \leqslant M, 0 \leqslant j \leqslant N, l_1 = Mh_1, l_2 = Nh_2\}$ avec frontière γ introduit dans le rectangle $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_2, \alpha = 1, 2\}$, sur lequel il s'agit de trouver la solution du *problème de Dirichlet au sens de différences finies pour l'équation de Poisson*

$$\begin{aligned} &y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} = -\varphi(x), \quad && x \in \omega, \\ &y(x) = g(x), && x \in \gamma, \end{aligned} \tag{3}$$

où

$$y_{\bar{x}_1 x_1} = \frac{1}{h_1^2}[y(i+1, j) - 2y(i, j) + y(i-1, j)],$$

$$y_{\bar{x}_2 x_2} = \frac{1}{h_2^2}[y(i, j+1) - 2y(i, j) + y(i, j-1)], \quad y(i, j) = y(x_{ij}).$$

Transformons le schéma (3). A cette fin multiplions (3) par $(-h_2^2)$ et répartissons entre les points la différence divisée $y_{\bar{x}_2 x_2}$. On a avec $1 \leqslant j \leqslant N-1$:

pour $2 \leqslant i \leqslant M-2$

$$-y(i, j-1) + [2y(i, j) - h_2^2 y_{\bar{x}_1 x_1}(i, j)] - y(i, j+1) = \\ = h_2^2 \varphi(i, j);$$

pour $i = 1$

$$-y(i, j-1) + \left[2y(i, j) - \frac{h_2^2}{h_1^2}(y(i+1, j) - 2y(i, j))\right] - \\ - y(i, j+1) = h_2^2 \bar{\varphi}(i, j);$$

pour $i = M-1$

$$-y(i, j-1) + \left[2y(i, j) - \frac{h_2^2}{h_1^2}(y(i-1, j) - 2y(i, j))\right] - \\ - y(i, j+1) = h_2^2 \bar{\varphi}(i, j),$$

où

$$\bar{\varphi}(1, j) = \varphi(1, j) + \frac{1}{h_1^2} g(0, j),$$

$$\bar{\varphi}(M-1, j) = \varphi(M-1, j) + \frac{1}{h_1^2} g(M, j).$$

De plus, pour $j = 0, N$, il vient

$$y(i, 0) = g(i, 0), \quad y(i, N) = g(i, N), \quad 1 \leqslant i \leqslant M-1.$$

Désignons maintenant par Y_j le vecteur de dimension $M-1$, dont les composantes sont les valeurs de la fonction de maille $y(i, j)$ aux nœuds intérieurs du maillage $\bar{\omega}$ sur la j-ième ligne:

$$Y_j = (y(1, j), y(2, j), \ldots, y(M-1, j)), \quad 0 \leqslant j \leqslant N,$$

et par $\boldsymbol{F}_j$ le vecteur de dimension $M-1$

$$\boldsymbol{F}_j = (h_2^2\overline{\varphi}(1, j),\ h_2^2\varphi(2, j), \ldots, h_2^2\varphi(M-2, j),\ h_2^2\overline{\varphi}(M-1, j)),$$
$$1 \leqslant j \leqslant N-1,$$

$$\boldsymbol{F}_j = (g(1, j),\ g(2, j), \ldots, g(M-1, j)),\ j = 0, N.$$

Définissons également la matrice carrée C de dimension $(M-1)\times \times (M-1)$ de la façon suivante:

$$CV = (\Lambda v(1),\ \Lambda v(2), \ldots, \Lambda v(M-1)),$$
$$V = (v(1),\ v(2), \ldots, v(M-1)),$$

où l'opérateur de différences Λ est

$$\Lambda v(i) = 2v(i) - h_2^2 v_{\bar{x}_1 x_1}(i), \quad 1 \leqslant i \leqslant M-1,$$
$$v(0) = v(M) = 0.$$

On voit sans peine que C est la matrice tridiagonale de la forme

$$C = \begin{Vmatrix} 2(1+\alpha) & -\alpha & 0 & \ldots & 0 & 0 & 0 \\ -\alpha & 2(1+\alpha) & -\alpha & \ldots & 0 & 0 & 0 \\ 0 & -\alpha & 2(1+\alpha) & \ldots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \ldots & 2(1+\alpha) & -\alpha & 0 \\ 0 & 0 & 0 & \ldots & -\alpha & 2(1+\alpha) & -\alpha \\ 0 & 0 & 0 & \ldots & 0 & -\alpha & 2(1+\alpha) \end{Vmatrix}, \tag{4}$$

où $\alpha = h_2^2/h_1^2$, C étant une matrice à dominance diagonale, car $|1+\alpha| > |\alpha|$, $\alpha > 0$, et, partant, n'est pas dégénérée.

Sur la base des notations introduites, les relations obtenues plus haut peuvent être écrites sous forme de système d'équations vectorielles triponctuelles

$$\begin{aligned} -\boldsymbol{Y}_{j-1} + C\boldsymbol{Y}_j - \boldsymbol{Y}_{j+1} &= \boldsymbol{F}_j, \quad 1 \leqslant j \leqslant N-1, \\ \boldsymbol{Y}_0 = \boldsymbol{F}_0, \quad \boldsymbol{Y}_N &= \boldsymbol{F}_N. \end{aligned} \tag{5}$$

C'est justement le système triponctuel cherché de forme spéciale à coefficients constants.

Le problème (5) est un cas particulier du problème général suivant: rechercher les vecteurs $\boldsymbol{Y}_j$ $(0 \leqslant j \leqslant N)$ satisfaisant au système qui suit:

$$\begin{aligned} &C_0\boldsymbol{Y}_0 - B_0\boldsymbol{Y}_1 = \boldsymbol{F}_0, && j = 0, \\ &-A_j\boldsymbol{Y}_{j-1} + C_j\boldsymbol{Y}_j - B_j\boldsymbol{Y}_{j+1} = \boldsymbol{F}_j, && 1 \leqslant j \leqslant N-1, \\ &-A_N\boldsymbol{Y}_{N-1} + C_N\boldsymbol{Y}_N = \boldsymbol{F}_N, && j = N, \end{aligned} \tag{6}$$

où Y_j et F_j sont des vecteurs de dimension M_j, C_j représente la matrice carrée $M_j \times M_j$, A_j et B_j sont les matrices rectangulaires de dimension $M_j \times M_{j-1}$ et $M_j \times M_{j+1}$ respectivement.

Aux systèmes de forme (6) se réduisent les schémas aux différences d'équations elliptiques d'ordre deux à coefficients variables dans un domaine arbitraire au nombre de dimensions quelconque. Au cas du domaine bidimensionnel, comme dans l'exemple examiné plus haut, le vecteur Y_j est formé par les inconnues de la j-ième ligne du maillage $\bar{\omega}$, tandis que au cas du domaine tridimensionnel il est formé par les inconnues de la j-ième couche du maillage $\bar{\omega}$. Dans ce dernier cas C_j sort des matrices tridiagonales par blocs avec matrices tridiagonales sur la diagonale principale.

Pour résoudre le système (6), on définira la *méthode du balayage matriciel* analogue à la méthode du balayage utilisée pour les équations triponctuelles scalaires.

2. Balayage des équations vectorielles triponctuelles. Construisons la méthode de résolution du système d'équations vectorielles triponctuelles (6). Ce système est apparenté au système d'équations scalaires triponctuelles, la méthode de résolution duquel a été étudiée au § 1. Aussi la solution du problème (6) sera-t-elle recherchée sous la forme

$$Y_j = \alpha_{j+1} Y_{j+1} + \beta_{j+1}, \quad j = N-1, \ N-2, \ \ldots, \ 0, \qquad (7)$$

où α_{j+1} est pour l'instant une matrice rectangulaire indéterminée de dimensions $M_j \times M_{j+1}$, et β_{j+1} le vecteur de dimension M_j. De la formule (7) et des équations du système (6) on tire pour $1 \leqslant j \leqslant N-1$ (comme pour un balayage ordinaire) les relations de récurrence servant au calcul des matrices α_j et des vecteurs β_j. A partir de (7) et des équations (6) pour $j = 0, N$, on obtient les valeurs initiales de α_1, β_1 et Y_N qui permettent de passer au calcul selon les relations de récurrence. Ecrivons les formules définitives de l'algorithme de la méthode proposée qui sera appelée *méthode du balayage matriciel*:

$$\alpha_{j+1} = (C_j - A_j\alpha_j)^{-1} B_j, \quad j = 1, 2, \ldots, N-1, \quad \alpha_1 = C_0^{-1} B_0, \qquad (8)$$

$$\beta_{j+1} = (C_j - A_j\alpha_j)^{-1} (F_j + A_j\beta_j), \quad j = 1, 2, \ldots, N, \quad \beta_1 = C_0^{-1} F_0, \qquad (9)$$

$$Y_j = \alpha_{j+1} Y_{j+1} + \beta_{j+1}, \quad j = N-1, N-2, \ldots, 0, \quad Y_N = \beta_{N+1}. \qquad (10)$$

On dira que l'algorithme (8)-(10) est *correct* si les matrices C_0 et $C_j - A_j\alpha_j$ ne sont pas dégénérées pour $1 \leqslant j \leqslant N$. Avant de passer à la définition de la stabilité de l'algorithme (8)-(10), rappelons quelques connaissances d'algèbre linéaire.

Soit A une matrice rectangulaire quelconque $m \times n$.

Soit $\| x \|_n$ la norme du vecteur x dans l'espace à n dimensions H_n. La norme de A se définit alors par l'égalité

$$\| A \| = \sup_{x \neq 0} \frac{\| Ax \|_m}{\| x \|_n}.$$

La norme A se définit apparemment par la matrice A, ainsi que par les normes vectorielles introduites dans H_n et H_m. Au cas de normes euclidiennes sur H_n et H_m ($\| x \|_n^2 = \sum_{i=1}^{n} x_i^2$), il vient $\| A \| = \sqrt{\rho}$, où ρ est la valeur propre maximale en module de la matrice A^*A.

De la définition de la norme il s'ensuit une relation évidente $\| Ax \|_m \leqslant \| A \| \, \| x \|_n$.

Soient données ensuite les matrices A et B de dimensions $m \times n$ et $n \times k$ respectivement. En introduisant sur H_m, H_k et H_n les normes vectorielles et en définissant à l'aide de ces dernières les normes des matrices A, B et AB, on obtient les inégalités $\| AB \| \leqslant \| A \| \, \| B \|$.

On dira que l'algorithme est *stable* si l'estimation $\| \alpha_j \| \leqslant 1$ est satisfaite pour $1 \leqslant j \leqslant N$ (on admet que dans les espaces de dimension finie H_{M_j}, auxquels appartiennent les vecteurs Y_j, est introduite une norme du même type, par exemple la norme euclidienne).

L e m m e 5. *Si C_j sont des matrices non dégénérées pour $0 \leqslant j \leqslant N$, tandis que A_j et B_j sont des matrices non nulles pour $1 \leqslant j \leqslant N - 1$, les conditions*

$$\| C_0^{-1} B_0 \| \leqslant 1, \quad \| C_N^{-1} A_N \| \leqslant 1, \quad \| C_j^{-1} A_j \| + \| C_j^{-1} B_j \| \leqslant 1, \quad 1 \leqslant j \leqslant N - 1,$$

étant satisfaites, avec au moins dans l'une des inégalités une inégalité stricte, alors l'algorithme de la méthode du balayage matriciel (8)-(10) *est stable et correct.*

Esquissons seulement l'étape principale de la démonstration du lemme, laissant au lecteur le soin de l'achever. On s'appuie dans la démonstration sur l'assertion connue: si pour la matrice carrée S on a l'estimation $\| S \| \leqslant q < 1$, il existe alors une matrice inverse à $E - S$ avec $\| (E - S)^{-1} \| \leqslant 1/(1 - q)$.

Supposons maintenant que $\| \alpha_j \| \leqslant 1$. De là et des conditions du lemme il vient

$$\| C_j^{-1} A_j \alpha_j \| \leqslant \| C_j^{-1} A_j \| \leqslant 1 - \| C_j^{-1} B_j \| < 1.$$

Vu que $C_j^{-1} A_j \alpha_j$ est une matrice carrée, il existe donc des matrices inverses à $E - C_j^{-1} A_j \alpha_j$ et à $C_j - A_j \alpha_j$ avec $\| (E - C_j^{-1} A_j \alpha_j)^{-1} \| \leqslant \leqslant 1/\| C_j^{-1} B_j \|$. De là et à partir de (8) on obtient immédiatement

$$\| \alpha_{j+1} \| \leqslant \| (E - C_j^{-1} A_j \alpha_j)^{-1} C_j^{-1} B_j \| \leqslant \leqslant \| (E - C_j^{-1} A_j \alpha_j)^{-1} \| \, \| C_j^{-1} B_j \| \leqslant 1.$$

La démonstration s'achève par induction.

Appliquons le lemme 5 au système d'équations vectorielles triponctuelles (5) obtenues à partir du problème de Dirichlet discret pour l'équation de Poisson sur un rectangle. Le système (5) est un cas particulier de (6), où $C_j = C$, $B_j = A_j = E$, $1 \leqslant j \leqslant N - 1$, $C_0 = C_N = E$, $B_0 = A_N = 0$, la matrice carrée C étant définie dans (4). Les conditions du lemme 5 prennent la forme $\| C^{-1} \| \leqslant 0.5$ pour l'exemple considéré. Au cas de norme euclidienne, en vertu de la symétrie de C, on a

$$\| C^{-1} \| = \max_k | \lambda_k (C^{-1}) | = \frac{1}{\min\limits_k | \lambda_k (C) |},$$

où $\lambda_k (C)$ est la valeur propre de la matrice C. De la définition de C il s'ensuit que $\lambda_k (C)$ est la valeur propre de l'opérateur Λ défini plus haut

$$\Lambda v (i) - \lambda_k v (i) = (2 - \lambda_k) v (i) - h_2^2 v_{\bar{x}_1 x_1} (i) = 0,$$
$$v (0) = v (M) = 0, \quad 1 \leqslant i \leqslant M - 1.$$

Ce problème se ramène par substitution $\lambda_k = 2 + h_2^2 \mu_k$ au problème de valeurs propres étudié au point 1, § 5, ch. I pour le cas d'un opérateur de différences du type le plus simple: $v_{\bar{x}_1 x_1} + \mu_k v = 0$, $1 \leqslant i \leqslant M - 1$, $v (0) = v (M) = 0$. Vu que ce problème a pour solution la solution égale à

$$\mu_k = \frac{4}{h_1^2} \sin^2 \frac{k \pi h_1}{2 l_1} > 0, \quad k = 1, 2, \ldots, M - 1,$$

on a $\lambda_k = \lambda_k (C) = 2 + h_2^2 \mu_k > 2$. La condition $\| C^{-1} \| \leqslant 0{,}5$ est donc vérifiée. L'algorithme (8)-(10) étant appliqué à la solution du système (5) est correct et stable.

Examinons maintenant la question du volume de l'information intermédiaire mémorisée, ainsi que celle du nombre d'opérations arithmétiques exigé pour l'obtention de l'algorithme (8)-(10), en posant, pour simplifier, que dans

le système (6) toutes les matrices sont carrées et possèdent les dimensions $M \times M$, tandis que tous les vecteurs Y_j et F_j sont de dimension M. Dans ce cas les coefficients de balayage α_j seront des matrices carrées de dimension $M \times M$ et les vecteurs β_j posséderont la dimension M.

Pour la mise en œuvre de (8)-(10) on doit mémoriser toutes les matrices α_j pour $1 \leqslant j \leqslant N$, tous les vecteurs β_j pour $1 \leqslant j \leqslant N+1$ et la matrice $(C_N - A_N\alpha_N)^{-1}$ servant au calcul de β_{N+1}. Les vecteurs β_j peuvent être placés à l'endroit réservé aux inconnues de Y_{j-1}. Mais pour mémoriser toutes les matrices α_j et les matrices $(C_N - A_N\alpha_N)^{-1}$, il faut retenir $M^2 (N + 1)$ éléments de ces matrices, car habituellement les matrices α_j sont pleines et asymétriques. Le volume de l'information supplémentaire à retenir est dans ce cas M fois supérieur au nombre total d'inconnues du problème qui vaut $M(N+1)$.

Apprécions maintenant le nombre d'opérations arithmétiques de l'algorithme (8)-(10), compte tenu du fait que pour la résolution de la série de problèmes (6) aux seconds membres F_j différents les matrices de balayage α_j et la matrice $(C_N - A_N\alpha_N)^{-1}$ peuvent être calculées une seule fois, tandis que les vecteurs β_j, et Y_j doivent être recalculés pour chaque nouveau problème.

En général les matrices $C_j - A_j\alpha_j$ sont pour tout j des matrices pleines. Aussi pour leur inversion faudra-t-il $O(M^3)$ opérations arithmétiques. Ensuite, la multiplication de $(C_j - A_j\alpha_j)^{-1}$ par la matrice B_j exigera au maximum $O(M^3)$ opérations. Aussi le calcul de α_{j+1} sur la base de α_j donné à l'aide de la formule (8) exigera-t-il $O(M^3)$ opérations arithmétiques. Le calcul de tous les α_j et de la matrice $(C_N - A_N\alpha_N)^{-1}$ exigera $O(M^3N)$ opérations.

Si la matrice A_j est pleine, le calcul de β_{j+1} avec β_j donné et $(C_j - A_j\alpha_j)^{-1}$ calculée exigera: $2M^2$ multiplications et $2M^2 - M$ additions. Si A_j est une matrice diagonale, ce nombre diminue alors et on aura besoin de $M^2 + M$ multiplications et de $2M^2 - M$ additions. Donc le calcul de β_j pour $2 \leqslant j \leqslant N+1$ exigera en tout $2M^2N$ multiplications et $(2M^2 - M)N$ additions. En y adjoignant les opérations effectuées pour le calcul de β_1 avec C_0^{-1} donné (M^2 multiplications et $M^2 - M$ additions) on obtient finalement $M^2(2N + 1)$ multiplications et $M^2(2N+1) - M(N+1)$ additions.

Pour obtenir tous les Y_j pour $0 \leqslant j \leqslant N-1$ avec Y_N donné, il faudra effectuer M^2N multiplications et M^2N additions. En résumé, pour le calcul de β_j et de Y_j il faut $M^2(3N+1)$ opérations de multiplication et $M^2(3N+1) - M(N+1)$ opérations d'addition. Si l'on s'abstient de distinguer ces opérations, on aura en tout $Q \approx 6M^2N$ opérations. C'est justement le nombre d'opérations arithmétiques qu'il faut effectuer pour obtenir la solution de chaque nouveau problème de la série. Pour résoudre un problème unique de la forme (6) nécessitant également le calcul des matrices de balayage α_j, il faut $Q = O(M^3N + M^2N)$ opérations.

Soit donnée une série de n problèmes de la forme (6). Il faut alors effectuer $Q_n = O(M^3N) + 6nM^2N$ opérations. Dans ce cas le nombre total d'inconnues dans la série s'élève à $nM(N+1)$. Il s'ensuit que pour la recherche d'une inconnue il faudra effectuer $q \approx O\left(\frac{M^2}{n}\right) + 6M$ opérations arithmétiques. Ainsi avec l'accroissement de n le nombre relatif d'opérations exigées pour le calcul d'une inconnue diminue en restant toutefois toujours plus grand que $6M$. C'est en quoi diffère pour l'essentiel la méthode du balayage matriciel de la méthode du balayage scalaire, où le nombre relatif d'opérations exigées pour le calcul d'une inconnue est une quantité finie indépendante du nombre d'inconnues.

3. Balayage d'équations vectorielles biponctuelles. Examinons maintenant la méthode de résolution des équations vectorielles biponctuelles

$$\begin{gathered} P_{i+1}V_{i+1} - Q_iV_i = F_{i+1}, \quad 0 \leqslant i \leqslant N-1, \\ P_0V_0 = F_0, \quad Q_NV_N = F_{N+1}, \end{gathered} \tag{11}$$

où le vecteur V_i est le vecteur de dimension M, P_{i+1} et Q_i des matrices carrées $M \times M$ pour $0 \leqslant i \leqslant N-1$, P_0 et Q_N des matrices rectangulaires de dimensions $M_1 \times M$ et $M_2 \times M$ respectivement, $M_1 + M_2 = M$. Le vecteur F_{i+1} a pour dimension M avec $0 \leqslant \leqslant i \leqslant N-1$, tandis que F_0 et F_{N+1} ont pour dimension M_1 et M_2 respectivement.

Réduisons d'abord le système (11) au système de la forme (6). A cette fin représentons les matrices composant (11) de la façon suivante:

$$P_0 = \| P_0^{11} \mid -P_0^{12} \|, \qquad Q_N = \| -Q_N^{21} \mid Q_N^{22} \|,$$
$$P_{i+1} = \left\| \frac{P_{i+1}^{11} \mid -P_{i+1}^{12}}{P_{i+1}^{21} \mid -P_{i+1}^{22}} \right\|, \quad Q_i = \left\| \frac{Q_i^{11} \mid -Q_i^{12}}{Q_i^{21} \mid -Q_i^{22}} \right\|, \tag{12}$$

où P_i^{kl} et Q_i^{kl} pour $0 \leqslant i \leqslant N$ sont des matrices de dimensions $M_k \times M_l$, $k, l = 1, 2$. En accord avec la représentation (12) posons

$$V_i = \begin{pmatrix} v_i^1 \\ v_i^2 \end{pmatrix}, \quad 0 \leqslant i \leqslant N, \quad F_{i+1} = \begin{pmatrix} f_{i+1}^1 \\ f_{i+1}^2 \end{pmatrix}, \quad 0 \leqslant i \leqslant N-1,$$
$$F_0 = f_0^1, \quad F_{N+1} = f_{N+1}, \tag{13}$$

où v_i^k et f_i^k sont des vecteurs de dimension M_k, $k = 1, 2$. En utilisant (12) et (13), écrivons le système (11) sous la forme suivante:

$$P_0^{11} v_0^1 - P_0^{12} v_0^2 = f_0^1,$$
$$\left.\begin{aligned} -Q_i^{11} v_i^1 + Q_i^{12} v_i^2 + P_{i+1}^{11} v_{i+1}^1 - P_{i+1}^{12} v_{i+1}^2 = f_{i+1}^1, \\ -Q_i^{21} v_i^1 + Q_i^{22} v_i^2 + P_{i+1}^{21} v_{i+1}^1 - P_{i+1}^{22} v_{i+1}^2 = f_{i+1}^2, \end{aligned}\right\} 0 \leqslant i \leqslant N-1, \tag{14}$$
$$-Q_N^{21} v_N^1 + Q_N^{22} v_N^2 = f_{N+1}^2.$$

Introduisons maintenant des nouveaux vecteurs d'inconnues en posant

$$Y_0 = v_0^1, \quad Y_{N+1} = v_N^2, \quad Y_{i+1} = \begin{pmatrix} v_i^2 \\ v_{i+1}^1 \end{pmatrix}, \quad 0 \leqslant i \leqslant N-1,$$

et des matrices

$$C_0 = P_0^{11}, \quad B_0 = \| P_0^{12} \mid 0^{11} \|, \quad C_{N+1} = Q_N^{22}, \quad A_{N+1} = \| 0^{22} \mid Q_N^{21} \|,$$
$$A_1 = \left\| \frac{Q_0^{11}}{Q_0^{21}} \right\|, \quad B_N = \left\| \frac{P_N^{12}}{P_N^{22}} \right\|, \quad A_{i+1} = \left\| \frac{0^{12} \mid Q_i^{11}}{0^{22} \mid Q_i^{21}} \right\|, \quad 1 \leqslant i \leqslant N-1,$$
$$B_{i+1} = \left\| \frac{P_{i+1}^{12} \mid 0^{11}}{P_{i+1}^{22} \mid 0^{21}} \right\|, \quad 0 \leqslant i \leqslant N-2, \quad C_{i+1} = \left\| \frac{Q_i^{12} \mid P_{i+1}^{11}}{Q_i^{22} \mid P_{i+1}^{21}} \right\|,$$
$$0 \leqslant i \leqslant N-1,$$

où 0^{kl} est la matrice nulle de dimensions $M_k \times M_l$, $k, l = 1, 2$.

Dans ces notations le système (14) prendra la forme

$$\begin{aligned} &C_0Y_0 - B_0Y_1 = F_0, && i = 0, \\ &-A_iY_{i-1} + C_iY_i - B_iY_{i+1} = F_i, && 1 \leqslant i \leqslant N, \\ &-A_{N+1}Y_N + C_{N+1}Y_{N+1} = F_{N+1}, && i = N + 1. \end{aligned} \qquad (15)$$

Bref, le système d'équations vectorielles biponctuelles (11) se réduit au système d'équations vectorielles triponctuelles de la forme (15) dont la méthode du balayage matriciel a été construite au point 2. L'algorithme du balayage matriciel pour (15) prend la forme suivante:

$$\alpha_{i+1} = (C_i - A_i\alpha_i)^{-1}B_i, \quad i = 1, 2, \ldots, N, \quad \alpha_1 = C_0^{-1}B_0. \qquad (16)$$

$$\beta_{i+1} = (C_i - A_i\alpha_i)^{-1}(F_i + A_i\beta_i), \quad i = 1, 2, \ldots, N + 1,$$

$$\beta_1 = C_0^{-1}F_0. \qquad (17)$$

$$Y_i = \alpha_{i+1}Y_{i+1} + \beta_{i+1}, \quad i = N, N - 1, \ldots, 0,$$

$$Y_{N+1} = \beta_{N+2}. \qquad (18)$$

de plus les matrices α_1 et α_{N+1} possèdent les dimensions $M_1 \times M$ et $M \times M_2$ respectivement, tandis que α_i est une matrice carrée de dimension $M \times M$ pour $2 \leqslant i \leqslant N$. Les vecteurs β_i ont la dimension M pour $2 \leqslant i \leqslant N + 1$, quant à β_1 et β_{N+2} leurs dimensions sont M_1 et M_2.

Transformons les formules (16)-(18). Compte tenu de la structure des matrices B_i on trouve que les matrices α_i ont la forme

$$\alpha_1 = \| \alpha_1^{12} \mid 0^{11} \|, \quad \alpha_{N+1} = \left\| \frac{\alpha_{N+1}^{22}}{\alpha_{N+1}^{12}} \right\|, \quad \alpha_i = \left\| \frac{\alpha_i^{22} \mid 0^{21}}{\alpha_i^{12} \mid 0^{11}} \right\|, \quad 2 \leqslant i \leqslant N. \qquad (19)$$

En portant (19) dans (16) et compte tenu de la définition des matrices A_i, B_i et C_i, on aboutit aux formules permettant de calculer α_i^{12} et α_i^{22}

$$\left\| \frac{\alpha_{i+1}^{22}}{\alpha_{i+1}^{12}} \right\| = \left\| \frac{Q_{i-1}^{12} - Q_{i-1}^{11}\alpha_i^{12} \mid P_i^{11}}{Q_{i-1}^{22} - Q_{i-1}^{21}\alpha_i^{12} \mid P_i^{21}} \right\|^{-1} \left\| \frac{P_i^{12}}{P_i^{22}} \right\|, \quad 1 \leqslant i \leqslant N, \qquad (20)$$

où $\alpha_1^{12} = (P_0^{11})^{-1}P_0^{12}$. Ensuite, en représentant le vecteur β_i sous la forme

$$\beta_1 = \beta_1^1, \quad \beta_{N+2} = \beta_{N+2}^2, \quad \beta_i = \begin{pmatrix} \beta_i^2 \\ \beta_i^1 \end{pmatrix}, \quad 2 \leqslant i \leqslant N + 1 \qquad (21)$$

et portant cette expression dans (17), il vient

$$\begin{pmatrix} \beta_{i+1}^2 \\ \beta_{i+1}^1 \end{pmatrix} = \left\| \begin{array}{c|c} Q_{i-1}^{12} - Q_{i-1}^{11}\alpha_i^{12} & P_i^{11} \\ \hline Q_{i-1}^{22} - Q_{i-1}^{21}\alpha_i^{12} & P_i^{21} \end{array} \right\|^{-1} \begin{pmatrix} f_i^1 + Q_{i-1}^{11}\beta_i^1 \\ f_i^2 + Q_{i-1}^{21}\beta_i^1 \end{pmatrix}, \quad 1 \leqslant i \leqslant N. \tag{22}$$

$$\beta_{N+2}^2 = \| Q_N^{22} - Q_N^{21}\alpha_{N+1}^{12} \|^{-1} (f_{N+1}^2 + Q_N^{21}\beta_{N+1}^1), \tag{23}$$

où $\beta_1^1 = \| P_0^{11} \|^{-1} f_0^1$.

Portons maintenant (19) et (21) dans (18) en utilisant les notations introduites pour Y_i. On obtient ainsi les formules suivantes permettant de calculer les composantes du vecteur d'inconnues:

$$\begin{aligned} v_{i-1}^2 &= \alpha_{i+1}^{22} v_i^2 + \beta_{i+1}^2, \quad i = N,\ N-1,\ \ldots,\ 1, \quad v_N^2 = \beta_{N+2}^2, \\ v_i^1 &= \alpha_{i+1}^{12} v_i^2 + \beta_{i+1}^1, \quad i = N,\ N-1,\ \ldots,\ 0. \end{aligned} \tag{24}$$

Bref, l'algorithme de la *méthode du balayage matriciel* pour le système d'équations vectorielles biponctuelles (11) se décrit par les formules (20), (22)-(24).

Vu que ces formules sont le corollaire de l'algorithme de balayage, utilisé à la résolution du système (15), auquel on a réduit le système initial d'équations vectorielles biponctuelles (11), les conditions suffisantes de correction et de stabilité de l'algorithme obtenu sont celles formulées dans le lemme 5 où il ne faut que substituer $N + 1$ à N, les matrices C_i, A_i et B_i étant définies plus haut.

En utilisant l'algorithme des balayages opposés pour le système (15), on est en mesure de construire l'algorithme correspondant au système initial d'équations vectorielles biponctuelles (11).

4. Balayage orthogonal pour équations vectorielles biponctuelles. Voyons encore une méthode de résolution du système d'équations biponctuelles (11) connue sous le nom de *méthode du balayage orthogonal.* Cette méthode comprend l'inversion des matrices P_i pour $1 \leqslant i \leqslant N$ et l'orthogonalisation des matrices orthogonales auxiliaires.

Recherchons la solution du système (11) sous la forme suivante:

$$V_i = B_i \beta_i + Y_i, \quad 0 \leqslant i \leqslant N, \tag{25}$$

où B_i pour tout i est une matrice rectangulaire de dimension $M \times M_2$, tandis que β_i et Y_i sont des vecteurs de dimensions M_2 et M respectivement.

En définissant B_0 et Y_0 à partir de la condition $P_0 B_0 = 0^{12}$, $P_0 Y_0 = F_0$, où 0^{12} est la matrice nulle de dimension $M_1 \times M_2$, on obtient que V_0 satisfait à la condition $P_0 V_0 = F_0$. Cherchons maintenant les formules de récurrence servant à la construction de proche en proche sur la base de B_0 et Y_0, des matrices B_i et des vecteurs Y_i.

Portons (25) dans (11). Si P_{i+1} est une matrice non dégénérée, il vient alors

$$B_{i+1}\beta_{i+1} + Y_{i+1} - P_{i+1}^{-1}Q_iB_i\beta_i = \\ = P_{i+1}^{-1}(F_{i+1} + Q_iY_i), \quad 0 \leqslant i \leqslant N-1,$$

ou

$$B_{i+1}\beta_{i+1} + Y_{i+1} - A_{i+1}\beta_i = X_{i+1}, \quad 0 \leqslant i \leqslant N-1, \tag{26}$$

où $A_{i+1} = P_{i+1}^{-1}Q_iB_i$, $X_{i+1} = P_{i+1}^{-1}(F_{i+1} + Q_iY_i)$. La matrice A_{i+1} a la dimension $M \times M_2$ et le vecteur X_{i+1} la dimension M.

Définissons B_{i+1} et Y_{i+1} de la façon suivante:

$$A_{i+1} = B_{i+1}\Omega_{i+1}, \quad Y_{i+1} = X_{i+1} - B_{i+1}\varphi_{i+1}, \tag{27}$$

où Ω_{i+1} et φ_{i+1} sont la matrice carrée $M_2 \times M_2$ et le vecteur de dimension M_2 encore non définis. Portant (27) dans (26) on obtient la relation $B_{i+1}(\beta_{i+1} - \Omega_{i+1}\beta_i) = B_{i+1}\varphi_{i+1}$, qui devient une identité si l'on pose

$$\Omega_{i+1}\beta_i = \beta_{i+1} - \varphi_{i+1}, \quad 0 \leqslant i \leqslant N-1. \tag{28}$$

Bref, si sont données des matrices Ω_i non dégénérées pour $1 \leqslant i \leqslant N$ et les vecteurs φ_i pour les mêmes i, on peut avec les formules (27) obtenir sur la base de B_0 et Y_0 toutes les matrices nécessaires B_i ainsi que les vecteurs Y_i pour $1 \leqslant i \leqslant N$.

Il reste à déterminer les vecteurs β_i. A partir de (25), pour $i = N$, et du système (11), on obtient deux relations $V_N = B_N\beta_N + Y_N$, $Q_NV_N = F_{N+1}$ avec B_N et Y_N connus. De là cherchons pour β_N l'équation $Q_NB_N\beta_N = F_{N+1} - Q_NY_N$ à la matrice carrée Q_NB_N de dimension $M_2 \times M_2$. Cette relation peut s'écrire sous forme de (28)

$$\Omega_{N+1}\beta_N = \beta_{N+1} - \varphi_{N+1}, \tag{29}$$

où $\beta_{N+1} = F_{N+1}$, $\varphi_{N+1} = Q_NY_N$, $\Omega_{N+1} = Q_NB_N$.

Si la matrice Ω_{N+1} n'est pas dégénérée, on trouve successivement à l'aide des formules (28), (29), en partant de β_{N+1}, tous les β_i pour $0 \leqslant i \leqslant N$. La solution du système (11) peut alors être obtenue au moyen des formules (25).

Vu l'arbitraire du choix de matrices Ω_i et de vecteurs φ_i, les formules mentionnées plus haut décrivent plutôt le principe de construction de la méthode de résolution du système (11) que l'algorithme concret. Un choix déterminé de Ω_i et de φ_i engendre une certaine méthode capable de résoudre (11). De telles méthodes seront toujours appelées balayages, qui dans le sens direct permettent de calculer B_i et Y_i et par remontée de trouver β_i ainsi que la solution V_i.

Arrêtons-nous maintenant sur un mode de choix de Ω_i et de φ_i. Vu que les formules (27) et (28) supposent l'inversion de la matrice Ω_{i+1}, cette dernière doit être inversible de manière suffisamment simple.

Dans la méthode de balayage orthogonal étudiée la matrice Ω_{i+1} et le vecteur φ_{i+1} sont impliqués par les exigences : 1) la matrice B_{i+1} se construit par orthonormalisation des colonnes de la matrice A_{i+1}; 2) le vecteur Y_{i+1} doit être orthogonal aux colonnes de la matrice B_{i+1}.

Ces exigences conduisent aux égalités

$$B^*_{i+1}B_{i+1} = E^{22}, \quad B^*_{i+1}Y_{i+1} = 0, \tag{29'}$$

où B^*_{i+1} est la matrice transposée à B_{i+1}, tandis que E^{22} est la matrice unité de dimension $M_2 \times M_2$.

Cherchons d'abord l'expression de φ_{i+1}. De (27) et de (29'), on tire $0 = B^*_{i+1}Y_{i+1} = B^*_{i+1}X_{i+1} - B^*_{i+1}B_{i+1}\varphi_{i+1} = B^*_{i+1}X_{i+1} - \varphi_{i+1}$. En résumé, le vecteur φ_{i+1} est défini : $\varphi_{i+1} = B^*_{i+1}X_{i+1}$.

Construisons maintenant les matrices Ω_{i+1} et B_{i+1}. Il y a plusieurs modes d'orthonormalisation des colonnes de la matrice A_{i+1}. On choisira l'orthonormalisation de Gram-Schmidt.

Soit la matrice A_{i+1} de rang M_2. Désignons par a_k et b_k les k-ièmes colonnes des matrices A_{i+1} et B_{i+1} respectivement et par (,) le produit scalaire des vecteurs. En guise de b_1 prenons la colonne normée a_1

$$b_1 = a_1/\omega_{11}, \quad \omega_{11} = \sqrt{(a_1, a_1)}. \tag{30}$$

Cherchons ensuite la colonne b_k sous la forme

$$b_k = \frac{1}{\omega_{kk}}\left(a_k - \sum_{n=1}^{k-1}{}' \omega_{nk}b_n\right), \quad 2 \leqslant k \leqslant M_2, \tag{31}$$

où les coefficients ω_{nk} s'obtiennent à partir de la condition d'orthogonalité du vecteur b_k aux vecteurs $b_1, b_2, \ldots, b_{k-1}$, et ω_{kk} de la condition de normalisation de b_k :

$$\omega_{nk} = (b_n, a_k), \quad n = 1, 2, \ldots, k-1, \quad \omega_{kk} = \sqrt{(a_k, a_k) - \sum_{n=1}^{k-1} \omega_{nk}^2}. \tag{32}$$

En vertu de l'hypothèse faite relativement au rang de A_{i+1}, les colonnes a_k sont linéairement indépendantes pour $1 \leqslant k \leqslant M_2$ et le processus d'orthonormalisation s'effectue sans singularités.

Il s'ensuit de (30)-(32) que les matrices A_{i+1} et B_{i+1} sont liées par la relation $A_{i+1} = B_{i+1}\Omega_{i+1}$, où Ω_{i+1} est la matrice carrée triangulaire supérieure de dimension $M_2 \times M_2$ à éléments ω_{nk} pour $1 \leqslant n \leqslant M_2$, $n \leqslant k \leqslant M_2$, définis dans (30) et (32) et $\omega_{nk} = 0$ pour $k < n$.

Bref, les formules (30)-(32) déterminent les matrices B_{i+1} et Ω_{i+1}. Un calcul élémentaire montre que la construction des matrices B_{i+1} et Ω_{i+1} peut être réalisée en effectuant $MM_2^2 + 0{,}5\,(M_2^2 - M_2)$ multiplications, $MM_2^2 - M_2$ additions et soustractions, MM_2 divisions et M_2 extractions de racine carrée. Le processus d'orthonormalisation doit être réalisé N fois dans le sens direct de la méthode du balayage. Cela exigera $O\,(MNM_2^2)$ opérations arithmétiques et NM_2 extractions de racine carrée.

Il nous reste à indiquer comment on obtient la matrice B_0 et le vecteur Y_0. Posons que les matrices P_{i+1} et Q_i sont non dégénérées pour $0 \leqslant i \leqslant N-1$. De plus, admettons que la matrice P_0^{11} est non dégénérée, quant à la matrice Q_N elle est de rang M_2.

Construisons B_0 et Y_0. Soit

$$A_0 = \left\| \frac{(P_0^{11})^{-1} P_0^{12}}{E^{22}} \right\|, \quad X_0 = \begin{pmatrix} (P_0^{11})^{-1} F_0 \\ 0 \end{pmatrix}$$

une matrice rectangulaire de dimension $M \times M_2$ et le vecteur de dimension M. Vu que la dimension de la matrice carrée unitaire E^{22} est $M_2 \times M_2$, le rang de A_0 vaut M_2. La matrice B_0 se construit sur la base de A_0 par orthonormalisation (30)-(32), tandis que Y_0 est choisi à l'aide de la formule $Y_0 = X_0 - B_0\varphi_0$ sur la base de la condition d'orthogonalité aux colonnes de la matrice B_0, ce qui donne $\varphi_0 = B_0^* X_0$. Vu que

$$B_0 = A_0\Omega_0^{-1}, \quad P_0A_0 = \| P_0^{11} \mid -P_0^{12} \| \left\| \frac{(P_0^{11})^{-1} P_0^{12}}{E^{22}} \right\| = \| 0^{12} \|,$$

$P_0B_0 = 0^{12}$. Ensuite, on a

$$P_0Y_0 = P_0X_0 - P_0B_0\varphi_0 = P_0X_0 = F_0.$$

B_0 et Y_0 ainsi construits vérifient donc les relations imposées: $P_0B_0 = 0^{12}$ et $P_0Y_0 = F_0$.

Remarquons que du fait de la non-dégénérescence de P_{i+1} et Q_i le rang de la matrice A_{i+1} coïncide avec celui de B_i. En outre, en vertu de la non-dégénérescence de Ω_0 le rang de B_0 coïncide avec celui de A_0 et vaut M_2. Aussi le processus d'orthonormalisation (30)-(32) s'effectuera-t-il sans entraves. Ensuite, puisque les rangs des matrices Q_N et B_N valent M_2, la matrice carrée $\Omega_{N+1} = Q_NB_N$ sera non dégénérée, permettant ainsi de trouver le vecteur β_N.

Donc l'algorithme de la méthode du balayage orthogonal prend la forme suivante:

1) $B_i\Omega_i = A_i, \quad i = 0, 1, 2, \ldots, N,$

$$A_i = P_i^{-1}Q_{i-1}B_{i-1}, \quad 1 \leqslant i \leqslant N, \quad A_0 = \left\| \frac{(P_0^{11})^{-1}P_0^{12}}{E^{22}} \right\|. \tag{33}$$

Les matrices B_i et Ω_i pour $0 \leqslant i \leqslant N$ se calculent à l'aide des formules (30)-(32) et sont mémorisées. On pose $\Omega_{N+1} = Q_N B_N$.

2) $Y_i = X_i - B_i\varphi_i, \quad \varphi_i = B_i^* X_i, \quad i = 0, 1, \ldots, N,$

$$X_i = P_i^{-1}(F_i + Q_{i-1}Y_{i-1}), \quad 1 \leqslant i \leqslant N, \quad X_0 = \begin{pmatrix} (P_0^{11})^{-1}F_0 \\ 0 \end{pmatrix}. \tag{34}$$

On calcule et on mémorise les vecteurs Y_i pour $0 \leqslant i \leqslant N$ et φ_i pour $1 \leqslant i \leqslant N$. On pose $\varphi_{N+1} = Q_N Y_N$.

3) $\Omega_{i+1}\beta_i = \beta_{i+1} - \varphi_{i+1}, \quad i = N, N-1, \ldots, 0,$

$$\beta_{N+1} = F_{N+1},$$

$$V_i = B_i\beta_i + Y_i, \quad 0 \leqslant i \leqslant N. \tag{35}$$

R e m a r q u e. Vu que les matrices Ω_i pour $1 \leqslant i \leqslant N$ sont des matrices triangulaires supérieures de dimension $M_2 \times M_2$, pour trouver β_i sur la base de β_{i+1} et φ_{i+1} il faut $O(M_2^2)$ opérations.

En guise d'illustration de l'algorithme proposé prenons un exemple. Supposons qu'il s'agit de résoudre le problème de différences triponctuel suivant:

$$\begin{aligned} &-y_{i-1} + y_i - y_{i+1} = 0, \quad 1 \leqslant i \leqslant N-1, \\ &y_0 = 1, \quad y_N = 0. \end{aligned} \tag{36}$$

Ce problème a été déjà vu au point 4 du § 2, où, au moyen de la méthode du balayage non monotone triponctuel, on a abouti à sa solution pour N non multiples de 3, à savoir

$$y_i = \frac{\sin\frac{(N-i)\pi}{3}}{\sin\frac{N\pi}{3}}, \quad 0 \leqslant i \leqslant N.$$

Réduisons le système (36) au système d'équations vectorielles biponctuelles de la forme (11) en posant

$$V_i = \begin{pmatrix} y_i \\ y_{i+1} \end{pmatrix}, \quad 0 \leqslant i \leqslant N-1.$$

On voit sans peine que (36) est équivalent au système suivant:

$$\begin{aligned} &V_{i+1} - QV_i = 0, \quad 0 \leqslant i \leqslant N-2, \\ &P_0V_0 = 1, \quad Q_{N-1}V_{N-1} = 0, \end{aligned} \tag{37}$$

Tableau 3

i	0	1	2	3	4	5	6	7	8	9	10	11
Ω_i	1	$\sqrt{2}$	$\frac{1}{\sqrt{2}}$	1	$\sqrt{2}$	$\frac{1}{\sqrt{2}}$	1	$\sqrt{2}$	$\frac{1}{\sqrt{2}}$	1	$\sqrt{2}$	$-\frac{1}{\sqrt{2}}$
φ_i	0	$-\frac{1}{\sqrt{2}}$	$-\frac{1}{2}$	1	$-\frac{1}{\sqrt{2}}$	$-\frac{1}{2}$	1	$-\frac{1}{\sqrt{2}}$	$-\frac{1}{2}$	1	$-\frac{1}{\sqrt{2}}$	$\frac{1}{2}$
β_i	1	$\frac{1}{\sqrt{2}}$	0	1	$\frac{1}{\sqrt{2}}$	0	1	$\frac{1}{\sqrt{2}}$	0	1	$\frac{1}{\sqrt{2}}$	0
B_i	$\begin{pmatrix} 0 \\ 1 \end{pmatrix}$	$\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ -1 \end{pmatrix}$	$\begin{pmatrix} -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$	$\begin{pmatrix} -1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \end{pmatrix}$	$\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ -1 \end{pmatrix}$	$\begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$	
Y_i	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}$	$\begin{pmatrix} 0 \\ -1 \end{pmatrix}$	$\begin{pmatrix} -1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -\frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}$	$\begin{pmatrix} 0 \\ -1 \end{pmatrix}$	$\begin{pmatrix} -1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -\frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$	
y_i	1	1	0	-1	-1	0	1	1	0	-1	-1	0

où $P_0 = \| 1 \mid 0 \|$, $Q_{N-1} = \| 0 \mid 1 \|$, $Q = \left\| \frac{0 \mid 1}{-1 \mid 1} \right\|$. Le système (37) est un cas particulier de (11) avec $M_1 = M_2 = 1$, $M = 2$.

Pour la résolution de (37) utilisons l'algorithme du balayage orthogonal (33)-(35). Dans l'exemple étudié les matrices B_i ont la dimension 2×1, Ω_i la dimension 1×1, les vecteurs Y_i auront la dimension 2 et les vecteurs β_i et φ_i la dimension 1.

On a donné au tableau 3 les matrices B_i et Ω_i, ainsi que les vecteurs Y_i, φ_i et β_i pour $N = 11$. La méthode du balayage orthogonal utilisée permet d'obtenir la solution exacte y_i du problème (36).

5. Balayage des équations triponctuelles à coefficients constants. Revenons de nouveau à la méthode du balayage matriciel des équations triponctuelles et examinons le cas particulier de ces équations, à savoir:

$$\begin{aligned} &-Y_{j-1} + CY_j - Y_{j+1} = F_j, \quad 1 \leqslant j \leqslant N-1, \\ &Y_0 = F_0, \quad Y_N = F_N, \end{aligned} \tag{38}$$

où C est la matrice carrée de dimension $M \times M$, tandis que Y_j et F_j sont les vecteurs cherché et donné de dimension M.

On a montré au point 1 qu'au système d'équations triponctuelles de la forme (38) se réduit le problème discret de Dirichlet pour l'équation de Poisson sur maillage rectangulaire donné dans un rectangle, la matrice C étant symétrique et tridiagonale. Ensuite, au point 2, § 4 on a montré que la méthode du balayage matriciel présentant pour (38) la forme

$$\alpha_{j+1} = (C - \alpha_j)^{-1}, \qquad j = 1, 2, \ldots, N-1, \quad \alpha_1 = 0. \tag{39}$$

$$\beta_{j+1} = \alpha_{j+1}(F_j + \beta_j), \qquad j = 1, 2, \ldots, N-1, \quad \beta_1 = F_0. \tag{40}$$

$$Y_j = \alpha_{j+1} Y_{j+1} + \beta_{j+1}, \quad j = N-1, N-2, \ldots, 1, \quad Y_N = F_N. \tag{41}$$

était correcte et stable. On y a également montré que les valeurs propres de la matrice C sont supérieures à 2:

$$\lambda_k = \lambda_k(C) = 2 + 4\frac{h_2^2}{h_1^2} \sin^2 \frac{k\pi h_1}{2l_1} > 2. \tag{42}$$

Rappelons que dans le cas d'équations vectorielles triponctuelles du type général, pour obtenir l'algorithme du balayage matriciel il faut $O(M^3N)$ opérations arithmétiques pour le calcul des matrices α_j et $O(M^2N)$ opérations pour le calcul des vecteurs de balayage β_j et l'obtention de la solution Y_j. Pour mémoriser les matrices pleines et, en général, non symétriques α_j il faut retenir $M^2(N+1)$ éléments de ces matrices. Diminue-t-on ce nombre si l'on résout

par la méthode du balayage matriciel le système vectoriel triponctuel spécial (38) à coefficients constants?

Pour l'exemple considéré, toutes les matrices α_j seront symétriques en vertu de la symétrie de la matrice C, or, bien que C soit une matrice tridiagonale, toutes les matrices α_j, $j \geqslant 2$, seront pleines. Donc on ne peut diminuer, compte tenu de la symétrie des matrices α_j, que le volume de l'information mémorisée intermédiaire et pas plus que de moitié. L'ordre du nombre d'opérations arithmétiques en M et N ne variera pas.

Construisons maintenant la modification de l'algorithme (39)-(41) qui ne nécessite pas de mémoire auxiliaire pour la mémorisation de l'information intermédiaire et se réalise avec $O(MN^2)$ opérations arithmétiques si le problème (38) est résolu avec la matrice tridiagonale C.

Cherchons d'abord la forme explicite des matrices de balayage α_j pour tout j. A cette fin, utilisant (39), exprimons α_j en fonction de la matrice C. Notant que

$$\alpha_1 = 0, \quad \alpha_2 = C^{-1}, \quad \alpha_3 = (C^2 - E)^{-1}C, \tag{43}$$

cherchons la solution de l'équation aux différences non linéaire (39) sous la forme

$$\alpha_j = P_{j-1}^{-1}(C)\, P_{j-2}(C), \quad j \geqslant 2, \tag{44}$$

où $P_j(C)$ est le polynôme en C de degré j. Récrivons (39) sous la forme

$$\alpha_{j+1}(C - \alpha_j) = E, \quad j \geqslant 2,$$

et portons-y (44). On obtient la relation de récurrence $P_j(C) = CP_{j-1}(C) - P_{j-2}(C)$, $j \geqslant 2$, ou, après le déplacement de l'indice d'une unité et compte tenu de (43):

$$\begin{gathered} P_{j+1}(C) = CP_j(C) - P_{j-1}(C), \quad j \geqslant 1, \\ P_0(C) = E, \quad P_1(C) = C. \end{gathered} \tag{45}$$

Bref, les formules (45) déterminent complètement le polynôme $P_j(C)$ pour tout $j \geqslant 0$.

Cherchons la solution de (45). Le polynôme algébrique correspondant vérifie complètement les relations

$$\begin{gathered} P_{j+1}(t) = tP_j(t) - P_{j-1}(t), \quad j \geqslant 1, \\ P_0(t) = 1, \quad P_1(t) = t, \end{gathered}$$

constitutant le problème de Cauchy pour l'équation aux différences triponctuelle à coefficients constants. Au point 2 du § 4, ch. I, on a trouvé la solution de ce problème $P_j(t) = U_j\left(\frac{t}{2}\right)$, $j \geqslant 0$,

où $U_j(x)$ est le polynôme de Tchébychev de seconde espèce de degré j

$$U_j(x) = \begin{cases} \dfrac{\sin((j+1)\arccos x)}{\sin \arccos x}, & |x| \leqslant 1, \\ \dfrac{\mathrm{sh}((j+1)\,\mathrm{Arch}\,x)}{\mathrm{sh}\,\mathrm{Arch}\,x}, & |x| \geqslant 1. \end{cases}$$

Donc l'expression explicite des matrices de balayage α_j est ainsi trouvée:

$$\alpha_j = U_{j-1}^{-1}\left(\frac{C}{2}\right) U_{j-2}\left(\frac{C}{2}\right), \quad j \geqslant 2, \quad \alpha_1 = 0. \tag{46}$$

Cela nous dispense d'effectuer des calculs, suivant la formule (39), des matrices de balayage α_j, qui constituent l'essentiel du volume des opérations arithmétiques de l'algorithme (39)-(41). En outre, il n'est pas nécessaire de mémoriser les matrices α_j.

Voyons maintenant les formules (40) et (41). Elles comprennent la multiplication de la matrice α_{j+1} par les vecteurs $F_j + \beta_j$ et Y_{j+1}. Montrons donc comment il est possible, sans le calcul de α_j à l'aide de la formule (46), d'obtenir le produit de la matrice α_j par un vecteur. Il faut pour cela recourir au lemme 6 qu'on formulera sans démonstration.

Lemme 6. *Soit un polynôme $f_n(x)$ de degré n possédant des racines simples. Le rapport du polynôme $g_m(x)$ de degré m au polynôme $f_n(x)$ de degré $n > m$ ne possédant pas de racines communes peut être représenté sous forme de somme de n fractions élémentaires*

$$\frac{g_m(x)}{f_n(x)} = \sum_{l=1}^{n} \frac{a_l}{x - x_l}, \quad a_l = \frac{g_m(x_l)}{f'_n(x_l)},$$

où x_l sont les racines de $f_n(x)$, et $f'_n(x)$ la dérivée du polynôme $f_n(x)$.

En utilisant le lemme 6, on obtient le développement en fractions simples du rapport $\varphi(x) = \dfrac{U_{j-2}(x)}{U_{j-1}(x)}$, $j \geqslant 2$. Vu que les racines de $U_{j-1}(x)$ sont

$$x_k = \cos\frac{k\pi}{j}, \quad k = 1, 2, \ldots, j-1,$$

et

$$U_{j-2}(x_k) = (-1)^{k-1}, \quad \frac{d}{dx}[U_{j-1}(x_k)] = \frac{j(-1)^{k-1}}{\sin^2\frac{k\pi}{j}},$$

en vertu du lemme 6 on a alors pour $\varphi(x)$ le développement suivant:

$$\varphi(x)=\frac{U_{j-2}(x)}{U_{j-1}(x)}=\sum_{k=1}^{j-1}\frac{\sin^2\frac{k\pi}{j}}{j}\left(x-\cos\frac{k\pi}{j}\right)^{-1}. \tag{47}$$

Il s'ensuit de (46) et (47) encore une représentation de la matrice α_j qui sera justement utilisée

$$\alpha_j=\sum_{k=1}^{j-1}a_{kj}\left(C-2\cos\frac{k\pi}{j}E\right)^{-1},\quad a_{kj}=\frac{2\sin^2\frac{k\pi}{j}}{j},\quad j\geqslant 2. \tag{48}$$

En utilisant (48), on peut réaliser la multiplication de la matrice α_j par le vecteur Y suivant l'algorithme: pour $k = 1, 2, \ldots, j-1$ on résout les équations

$$\left(C-2\cos\frac{k\pi}{j}E\right)V_k=a_{kj}Y, \tag{49}$$

où a_{kj} est défini dans (48), tandis que le résultat α_jY s'obtient de proche en proche par sommation des vecteurs V_k

$$\alpha_jY=\sum_{k=1}^{j-1}V_k. \tag{50}$$

Notons qu'en vertu de (42) la matrice $C-2\cos\frac{k\pi}{j}E$ est non dégénérée et, de plus, tridiagonale, si la matrice C l'était. Dans ce cas chacune des équations (49) se résout en $O(M)$ opérations arithmétiques à l'aide de la méthode du balayage triponctuel scalaire décrite au § 1. Donc la résolution de tous les problèmes (49) et le calcul de la somme (50) exigent $O(Mj)$ opérations. Comme dans (40) et (41) la multiplication de la matrice α_j par les vecteurs est effectuée pour $j = 2, 3, \ldots, N$, la méthode modifiée du balayage matriciel (40), (41) et (49), (50) vaut $O(MN^2)$ opérations arithmétiques.

On a donc construit la *méthode modifiée du balayage matriciel* qui permet de trouver la solution du problème discret de Dirichlet pour l'équation de Poisson dans le rectangle au moyen de $O(MN^2)$ opérations arithmétiques. La diminution du nombre d'opérations par rapport à celui exigé par l'algorithme initial (39)-(41) est le résultat de la prise en compte de la nature spécifique du problème à résoudre.

On étudiera dans les deux chapitres suivants d'autres méthodes directes de résolution du problème mentionné, ainsi que des problèmes de différences semblables qui exigeront un nombre encore plus faible d'opérations que la méthode construite ici.

CHAPITRE III

MÉTHODE DE RÉDUCTION TOTALE

On étudie dans ce chapitre la méthode de résolution des équations de mailles elliptiques spéciales sur maillage ou méthode de réduction totale. Cette méthode directe permet de trouver la solution du problème discret de Dirichlet pour l'équation de Poisson dans un rectangle en $O(N^2 \log_2 N)$ opérations arithmétiques, où N est le nombre de nœuds du maillage suivant chaque direction.

Au § 1 sont définis les problèmes aux limites pour équations aux différence-dont la résolution peut être effectuée par la méthode de réduction. Au § 2 ess décrit l'algorithme de la méthode pour le cas du premier problème aux limitet et au § 3 sont étudiés des exemples d'application de la méthode. Au § 4 on a génés ralisé la méthode aux cas de conditions aux limites générales.

§ 1. Rroblèmes aux limites pour les équations vectorielles triponctuelles

1. Position des problèmes aux limites. Au chapitre II. pour résoudre les équations triponctuelles vectorielles et scalaires. on a construit les méthodes des balayages scalaire et matriciel. La méthode du balayage matriciel pour équations à coefficients variables est mise en œuvre avec $O(M^3N)$ opérations arithmétiques, où N est le nombre d'équations et M la dimension des vecteurs d'inconnues (le nombre d'inconnues dans le problème est égal à MN). Pour les classes spéciales d'équations vectorielles correspondant. par exemple. au problème discret de Dirichlet pour l'équation de Poisson dans un rectangle on a proposé l'algorithme modifié de la méthode du balayage matriciel. Cet algorithme permet de réduire le nombre d'opérations à $O(MN^2)$.

Le présent chapitre est consacré à l'étude subséquente des méthodes directes de résolution d'équations vectorielles de type spécial auxquelles se réduisent les schémas aux différences construits pour les plus simples équations elliptiques. On construira la *méthode de réduction totale* qui permet de résoudre les principaux problèmes aux limites en $O(MN \log_2 N)$ opérations arithmétiques. Si l'on ne tient pas compte de la faible dépendance logarithmique de N. le nombre d'opérations exigé par la méthode est proportionnel au nombre d'inconnues MN. L'élaboration de cette méthode constitue

un apport substantiel au développement des méthodes directes et itératives de résolution des équations de mailles.

Formulons les problèmes aux limites pour les équations vectorielles triponctuelles dont la solution peut être obtenue par la méthode de réduction totale. On examinera les problèmes suivants:

1) *Premier problème aux limites.* Trouver la solution de l'équation

$$-Y_{j-1} + CY_j - Y_{j+1} = F_j, \quad 1 \leqslant j \leqslant N-1, \tag{1}$$

satisfaisant aux valeurs données pour $j = 0$ et $j = N$

$$Y_0 = F_0, \quad Y_N = F_N. \tag{2}$$

Y_j est ici le vecteur d'inconnues de numéro j, F_j la partie droite donnée et C la matrice carrée donnée.

2) *Deuxième et troisième problèmes aux limites.* On cherche la solution de l'équation (1) qui satisfait aux conditions aux limites suivantes pour $j = 0$ et $j = N$:

$$\begin{aligned} (C + 2\alpha E)\, Y_0 - 2Y_1 &= F_0, \quad j = 0, \\ -2Y_{N-1} + (C + 2\beta E)\, Y_N &= F_N, \quad j = N, \end{aligned} \tag{3}$$

où $\alpha \geqslant 0$, $\beta \geqslant 0$. Pour $\alpha = \beta = 0$, la formule (3) définit les conditions aux limites de seconde espèce. On examinera également les combinaisons des conditions aux limites, par exemple, si pour $j = 0$ ce sont les conditions aux limites de première espèce qui sont définies, tandis que pour $j = N$ sont définies les conditions de deuxième et troisième espèces.

3) *Problème aux limites périodique.* Trouver la solution de l'équation $-Y_{j-1} + CY_j - Y_{j+1} = F_j$ constituant une équation périodique, $Y_{N+j} = Y_j$. On suppose que le second membre F_j est également périodique, $F_{N+j} = F_j$. Ce problème se formule de la façon équivalente suivante: trouver la solution de l'équation

$$\begin{aligned} &-Y_{j-1} + CY_j - Y_{j+1} = F_j, \quad 1 \leqslant j \leqslant N-1, \\ &-Y_{N-1} + CY_0 - Y_1 = F_0, \quad Y_N = Y_0. \end{aligned} \tag{4}$$

A cette espèce d'équations se réduisent les schémas aux différences pour les équations elliptiques en systèmes de coordonnées curvilignes orthogonales: cylindriques, polaires et sphériques.

Outre l'équation vectorielle de base (1) contenant la seule matrice C, on examinera le premier problème aux limites pour une équation plus générale

$$\begin{aligned} &-BY_{j-1} + AY_j - BY_{j+1} = F_j, \quad 1 \leqslant j \leqslant N-1, \\ &Y_0 = F_0, \quad Y_N = F_N \end{aligned} \tag{5}$$

aux matrices carrées A et B. Ces problèmes apparaissent avec la résolution du problème discret de Dirichlet de grande précision pour l'équation de Poisson dans un rectangle.

Formulons les exigences envers les matrices C, A et B qui garantissent l'applicabilité de la méthode de réduction totale à la résolution des problèmes (1)-(5). Admettons pour les problèmes (1)-(4) que pour tout vecteur Y se vérifie l'inégalité $(CY, Y) \geqslant 2\,(Y, Y)$, et pour le problème (5) l'inégalité $(AY, Y) \geqslant 2\,(BY, Y) > 0$. On utilise ici le produit scalaire trivial des vecteurs.

2. Premier problème aux limites. Commençons l'étude de la méthode de réduction totale par la description des problèmes aux limites discrets pour équations elliptiques qui peuvent être écrites sous forme d'équations vectorielles spéciales (1)-(5). Soit un maillage carré $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leqslant i \leqslant M, 0 \leqslant j \leqslant N, h_1 = l_1/M, h_2 = l_2/N)\}$ avec frontière γ, introduit dans le rectangle $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha, \alpha = 1, 2\}$, sur lequel il s'agit de trouver la solution du problème discret de Dirichlet pour l'équation de Poisson

$$\begin{aligned} y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} &= -\varphi(x), \quad x \in \omega, \\ y(x) &= g(x), \quad x \in \gamma. \end{aligned} \tag{6}$$

Au § 4, ch. II, on a montré que le problème (6) peut être écrit sous forme de (1), (2), où Y_j est le vecteur de dimension $M-1$ et dont les composantes constituent des valeurs de la fonction de maille $y(i, j) = y(x_{ij})$ aux nœuds internes de la j-ième ligne du maillage $\bar{\omega}$:

$$Y_j = (y(1, j), y(2, j), \ldots, y(M-1, j)), \quad 0 \leqslant j \leqslant N.$$

C est la matrice carrée de dimension $(M-1) \times (M-1)$ correspondant à l'opérateur de différences Λ, où

$$\begin{aligned} \Lambda y &= 2y - h_2^2 y_{\bar{x}_1 x_1}, \quad h_1 \leqslant x_1 \leqslant l_1 - h_1, \\ y &= 0, \quad x_1 = 0, \ l_1. \end{aligned} \tag{7}$$

Le second membre F_j est le vecteur de dimension $M-1$ qui se définit de la façon suivante:

1) pour $j = 1, 2, \ldots, N-1$

$$F_j = (h_2^2\bar{\varphi}(1, j), h_2^2\varphi(2, j), \ldots, h_2^2\varphi(M-2, j), h_2^2\bar{\varphi}(M-1, j)), \tag{8}$$

où

$$\bar{\varphi}(1, j) = \varphi(1, j) + \frac{1}{h_1^2} g(0, j),$$

$$\bar{\varphi}(M-1, j) = \varphi(M-1, j) + \frac{1}{h_1^2} g(M, j);$$

2) pour $j = 0, N$

$$F_j = (g(1, j), g(2, j), \ldots, g(M-1, j)). \qquad (9)$$

Il s'ensuit de (7) que pour l'exemple étudié la matrice C est une matrice tridiagonale symétrique.

Examinons un problème de différences plus compliqué qui s'écrit également sous forme des équations (1), (2). Sur le maillage $\bar{\omega}$ il s'agit de trouver la solution de l'équation aux différences de Poisson

$$y_{\bar{x}_1x_1} + y_{\bar{x}_2x_2} = -\varphi(x), \quad x \in \omega, \qquad (10)$$

satisfaisant sur les côtés $x_1 = 0$ et $x_1 = l_1$ aux conditions aux limites de troisième ou de seconde espèce

$$\frac{2}{h_1} y_{x_1} + y_{\bar{x}_2x_2} = \frac{2}{h_1} \varkappa_{-1} y - \bar{\varphi}, \quad x_1 = 0, \qquad (11)$$

$$-\frac{2}{h_1} y_{\bar{x}_1} + y_{\bar{x}_2x_2} = \frac{2}{h_1} \varkappa_{+1} y - \bar{\varphi}, \quad x_1 = l_1, \qquad (12)$$

$$h_2 \leqslant x_2 \leqslant l_2 - h_2$$

et aux conditions aux limites de première espèce sur les côtés $x_2 = 0$, $x_2 = l_2$: $y(x) = g(x)$, $x_2 = 0, l_2$, $0 \leqslant x_1 \leqslant l_1$. Pour que le problème posé puisse être écrit sous forme de (1), (2) avec la matrice C indépendante de j, il faut supposer remplie la condition $\varkappa_{\pm 1} = \text{const}$.

Réduisons ce problème à (1), (2). A cette fin multiplions (10)-(12) par $(-h_2^2)$ et répartissons la différence divisée $y_{\bar{x}_2x_2}$ entre les points pour tous les $j = 1, 2, \ldots, N-1$. On obtient les équations suivantes:

1) pour $i = 0$

$$-y(0, j-1) + 2\left[\left(1 + \frac{h_2^2}{h_1}\varkappa_{-1}\right) y(0, j) - \frac{h_2^2}{h_1} y_{x_1}(0, j)\right] - \\ - y(0, j+1) = h_2^2\bar{\varphi}(0, j);$$

2) pour $i = 1, 2, \ldots, M-1$

$$-y(i, j-1) + [2y(i, j) - h_2^2 y_{\bar{x}_1x_1}(i, j)] - y(i, j+1) = h_2^2\varphi(i, j);$$

3) pour $i = M$

$$-y(M, j-1) + 2\left[\left(1 + \frac{h_2^2}{h_1}\varkappa_{+1}\right) y(M, j) + \frac{h_2^2}{h_1} y_{\bar{x}_1}(M, j)\right] - \\ - y(M, j+1) = h_2^2\bar{\varphi}(M, j).$$

Posons

$$Y_j = (y(0, j), y(1, j), \ldots, y(M, j)), \quad 0 \leqslant j \leqslant N,$$

$$F_j = (h_2^2\bar{\varphi}(0, j), h_2^2\varphi(1, j), \ldots, h_2^2\varphi(M-1, j), h_2^2\bar{\varphi}(M, j)), \qquad (13)$$
$$1 \leqslant j \leqslant N-1,$$

$$F_j = (g(0, j), g(1, j), \ldots, g(M, j)), \quad j = 0, N.$$

Dans ces notations les équations obtenues s'écrivent en forme de (1), (2), où la matrice carrée C de dimension $(M+1)\times(M+1)$ correspond à l'opérateur de différences Λ:

$$\Lambda y=\begin{cases} 2\left(1+\frac{h_2^2}{h_1}\varkappa_{-1}\right)y-\frac{2h_2^2}{h_1}y_{x_1}, & x_1=0,\\ 2y-h_2^2y_{\bar{x}_1x_1}, & h_1\leqslant x_1\leqslant l_1-h_1,\\ 2\left(1+\frac{h_2^2}{h_1}\varkappa_{+1}\right)y+\frac{2h_2^2}{h_1}y_{\bar{x}_1}, & x_1=l_1.\end{cases}\tag{14}$$

On a ici de nouveau affaire au cas où C est une matrice tridiagonale. En posant sur les côtés $x_1=0$, l_1 les conditions aux limites de troisième espèce (11), (12) au lieu des conditions de première espèce, on n'aboutit qu'à une autre définition de l'opérateur Λ: à la place de (7) on a (14). L'aspect des équations (1) et des conditions aux limites (2) dans ce cas ne varie pas. Si pour $x_1=0$ au lieu de la condition (11) on a la condition aux limites de première espèce $y(x)=g(x)$, pour $x_1=l_1$ la condition (12) restant la même, le problème de différences ainsi posé se réduit aussi à (1), (2). Dans ce cas

$$Y_j=(y(1,j),\ y(2,j),\ \ldots,\ y(M,j)),\quad 0\leqslant j\leqslant N,$$

$$F_j=(h_2^2\bar{\varphi}(1,j),\ h_2^2\varphi(2,j),\ \ldots,\ h_2^2\varphi(M-1,j),\ h_2^2\bar{\varphi}(M,j)),$$
$$1\leqslant j\leqslant N-1,$$

où $\bar{\varphi}(1,j)=\varphi(1,j)+\frac{1}{h_1^2}g(0,j)$, $\bar{\varphi}(M,j)$ sont les valeurs au point correspondant du second membre de $\bar{\varphi}$ défini dans (12), tandis que la matrice carrée C correspond à l'opérateur de différences Λ, où

$$\Lambda y=\begin{cases} 2y-h_2^2y_{\bar{x}_1x_1}, & h_1\leqslant x_1\leqslant l_1-h_1,\\ 2\left(1+\frac{h_2^2}{k_1}\varkappa_{+1}\right)y+\frac{2h_2^2}{h_1}y_{\bar{x}_1}, & x_1=l_1\end{cases}\tag{15}$$

et $y=0$ pour $x_1=0$.

Si la condition aux limites de première espèce est posée pour $x_1=l_1$, tandis que la condition aux limites de troisième espèce (11) l'est pour $x_1=0$, on a alors dans (1), (2)

$$Y_j=(y(0,j),\ y(1,j),\ \ldots,\ y(M-1,j)),\quad 0\leqslant j\leqslant N,$$

$$F_j=(h_2^2\bar{\varphi}(0,j),\ h_2^2\varphi(1,j),\ \ldots,\ h_2^2\varphi(M-2,j),\ h_2^2\bar{\varphi}(M-1,j)),$$
$$1\leqslant j\leqslant N-1,$$

où $\bar{\varphi}(M-1, j) = \varphi(M-1, j) + \frac{1}{h_1^2} g(M, j)$, tandis que la matrice C correspond à l'opérateur de différences Λ, où

$$\Lambda y = \begin{cases} 2\left(1 + \frac{h_2^2}{h_1} \varkappa_{-1}\right) y - \frac{2h_2^2}{h_1} y_{x_1}, & x_1 = 0, \\ 2y - h_2^2 y_{\bar{x}_1 x_1}, & h_1 \leqslant x_1 \leqslant l_1 - h_1 \end{cases} \tag{16}$$

et $y = 0$ pour $x_1 = l_1$.

Bref, on a montré que si en direction de x_2 sont données les conditions aux limites de première espèce et en direction de x_1 une combinaison quelconque des conditions aux limites de première, de seconde ou de troisième espèce, les schémas aux différences pour l'équation de Poisson dans le rectangle s'écriront alors sous forme du premier problème aux limites pour les équations vectorielles triponctuelles (1), (2). La matrice C est déterminée à l'aide de l'opérateur de différences Λ qui, selon le type des conditions aux limites sur les côtés $x_1 = 0$ et $x_1 = l_1$, est donné par les formules (7), (14)-(16).

3. Autres problèmes aux limites pour équations aux différences. Le type des conditions aux limites pour l'équation (1) se détermine complètement par celui de l'équation aux différences (10) sur les côtés du rectangle $x_2 = 0$ et $x_2 = l_2$. On a étudié le cas quand sur ces côtés ont été données les conditions aux limites de première espèce.

Voyons maintenant d'autres problèmes aux limites de l'équation (10) qui se réduisent aux équations vectorielles (1), (3). Supposons que sur le rectangle du maillage $\bar{\omega}$, défini plus haut, il s'agit de trouver la solution du *troisième problème aux limites* pour l'équation aux différences de Poisson. Le schéma aux différences prend la forme suivante:

$$y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} = -\varphi(x), \qquad x \in \omega, \tag{17}$$

$$\frac{2}{h_1} y_{x_1} + y_{\bar{x}_2 x_2} = \frac{2}{h_1} \varkappa_{-1} y - \bar{\varphi}, \qquad x_1 = 0, \tag{18}$$

$$-\frac{2}{h_1} y_{\bar{x}_1} + y_{\bar{x}_2 x_2} = \frac{2}{h_1} \varkappa_{+1} y - \bar{\varphi}, \qquad x_1 = l_1, \quad h_2 \leqslant x_2 \leqslant l_2 - h_2.$$

$$y_{\bar{x}_1 x_1} + \frac{2}{h_2} y_{x_2} = \frac{2}{h_2} \varkappa_{-2} y - \bar{\varphi}, \qquad x_2 = 0, \tag{19}$$

$$y_{\bar{x}_1 x_1} - \frac{2}{h_2} y_{\bar{x}_2} = \frac{2}{h_2} \varkappa_{+2} y - \bar{\varphi}, \qquad x_2 = l_2, \quad h_1 \leqslant x_1 \leqslant l_1 - h_1. \tag{20}$$

L'approximation aux coins du maillage a une forme spéciale:

$$\frac{2}{h_1} y_{x_1} + \frac{2}{h_2} y_{x_2} = \left(\frac{2}{h_1} \varkappa_{-1} + \frac{2}{h_2} \varkappa_{-2}\right) y - \bar{\varphi}, \quad x_1 = 0, \quad x_2 = 0, \quad (21)$$

$$-\frac{2}{h_1} y_{\bar{x}_1} + \frac{2}{h_2} y_{x_2} = \left(\frac{2}{h_1} \varkappa_{+1} + \frac{2}{h_2} \varkappa_{-2}\right) y - \bar{\varphi}, \quad x_1 = l_1, \quad x_2 = 0, \quad (22)$$

$$\frac{2}{h_1} y_{x_1} - \frac{2}{h_2} y_{\bar{x}_2} = \left(\frac{2}{h_1} \varkappa_{-1} + \frac{2}{h_2} \varkappa_{+2}\right) y - \bar{\varphi}, \quad x_1 = 0, \quad x_2 = l_2, \quad (23)$$

$$-\frac{2}{h_1} y_{\bar{x}_1} - \frac{2}{h_2} y_{\bar{x}_2} = \left(\frac{2}{h_1} \varkappa_{+1} + \frac{2}{h_2} \varkappa_{+2}\right) y - \bar{\varphi}, \quad x_1 = l_1, \quad x_2 = l_2. \quad (24)$$

On admet ici que les conditions $\varkappa_{\pm\alpha} = \text{const}$, $\alpha = 1, 2$ sont satisfaites.

Montrons que le problème (17)-(24) se réduit à (1), (3). En effet, en désignant par Y_j le vecteur de dimension $M + 1$

$$Y_j = (y(0, j), y(1, j), \ldots, y(M, j)), \quad 0 \leqslant j \leqslant N$$

et en définissant le second membre F_j pour $j = 1, 2, \ldots, N - 1$ selon les formules (13), on obtient à partir de (17) et (18), comme au point précédent, l'équation (1) avec matrice C correspondant à Λ de (14). Il reste à montrer que les conditions (19)-(24) peuvent être écrites sous forme de conditions aux limites (3).

Multiplions (19), (21) et (22) par $(-h_2^2)$ et répartissons la différence divisée (y_{x_2}) qui y figure entre les points. Il vient

1) pour $i = 0$

$$2\left[\left(1 + \frac{h_2^2}{h_1} \varkappa_{-1}\right) y(0, 0) - \frac{h_2^2}{h_1} y_{x_1}(0, 0)\right] + \\ + 2h_2\varkappa_{-2} y(0, 0) - 2y(0, 1) = h_2^2\bar{\varphi}(0, 0);$$

2) pour $i = 1, 2, \ldots, M - 1$

$$[2y(i, 0) - h_2^2 y_{\bar{x}_1 x_1}(i, 0)] + 2h_2\varkappa_{-2} y(i, 0) - 2y(i, 1) = h_2^2\bar{\varphi}(i, 0);$$

3) pour $i = M$

$$2\left[\left(1 + \frac{h_2^2}{h_1} \varkappa_{+1}\right) y(M, 0) + \frac{h_2^2}{h_1} y_{\bar{x}_1}(M, 0)\right] + \\ + 2h_2\varkappa_{-2} y(M, 0) - 2y(M, 1) = h_2^2\bar{\varphi}(M, 0).$$

Si l'on pose $\alpha = h_2\varkappa_{-2}$, ces égalités peuvent être écrites sous forme vectorielle

$$(C + 2\alpha E) Y_0 - 2Y_1 = F_0, \quad (25)$$

où $F_0 = (h_2^2\bar{\varphi}(0, 0), h_2^2\bar{\varphi}(1, 0), \ldots, h_2^2\bar{\varphi}(M, 0))$.

De façon analogue, à partir de (20), (23) et (24), on obtient l'équation

$$-2Y_{N-1} + (C + 2\beta E) Y_N = F_N,$$

où on a posé $\beta = h_2\varkappa_{+2}$ et $F_N = (h_2^2\bar{\varphi}(0, N),\ h_2^2\bar{\varphi}(1, N), \ldots, h_2^2\bar{\varphi}(M, N))$. En résumé, le schéma aux différences (17)-(24) est réduit au problème (1), (3).

Examinons maintenant le cas où sont données quelques combinaisons de conditions aux limites sur les côtés du rectangle $\bar{G}$. Comme il a été montré plus haut, l'attribution aux côtés $x_1 = 0$ et $x_1 = l_1$ de conditions aux limites autres que (18) n'influe que sur la définition de la matrice C. Si pour $x_2 = 0$ est imposée la condition aux limites de première espèce, c'est-à-dire si (19), (21) et (22) sont remplacés par $y(x) = g(x)$, $x_2 = 0$, on doit alors substituer à (25) la condition $Y_0 = F_0$, où $F_0 = (g(0, 0), \ldots, g(M, 0))$. Dans ce cas le problème aux limites vectoriel triponctuel prend la forme

$$\begin{gathered} -Y_{j-1} + CY_j - Y_{j+1} = F_j, \quad 1 \leqslant j \leqslant N-1, \\ Y_0 = F_0, \\ -2Y_{N-1} + (C + 2\beta E)\, Y_N = F_N. \end{gathered} \tag{26}$$

On aboutit à un système analogue quand sur le côté $x_2 = l_2$ est imposée la condition aux limites de première espèce et sur le côté $x_2 = 0$ la condition aux limites de troisième espèce. Dans ce cas le problème aux limites vectoriel prend la forme

$$\begin{gathered} -Y_{j-1} + CY_j - Y_{j+1} = F_j, \quad 1 \leqslant j \leqslant N-1, \\ (C + 2\alpha E)\, Y_0 - 2Y_1 = F_0, \quad Y_N = F_N. \end{gathered} \tag{27}$$

On a passé en revue les exemples de problèmes aux limites pour l'équation aux différences de Poisson dans un rectangle et montré qu'il leur correspond les problèmes aux limites vectoriels (1), (2) ou (1), (3), ou (26), (27) à matrice tridiagonale C correspondante.

Aux problèmes aux limites vectoriels mentionnés se réduisent également les schémas aux différences d'équations elliptiques plus compliquées aussi bien en coordonnées cartésiennes que curvilignes orthogonales. Donnons des exemples. Dans le système cartésien ce sont les problèmes aux limites principaux des équations elliptiques

$$\frac{\partial}{\partial x_1}\left(k_1(x_1)\frac{\partial u}{\partial x_1}\right) + k_2(x_1)\frac{\partial^2 u}{\partial x_2^2} - q(x_1)\, u = -f(x), \quad x \in G,$$

dont les coefficients ne dépendent que d'une variable. Dans ce cas dans le rectangle $\bar{G}$ on peut introduire le maillage rectangulaire $\bar{\omega}$ au pas h_2 uniforme en direction de x_2 et aux pas quelconques non uniformes en direction de x_1.

Dans le système de coordonnées cylindrique les exemples nous sont fournis par les problèmes aux limites pour l'équation de Poisson dans un cylindre de révolution fini ou un tube au cas où a lieu une

symétrie axiale:

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right)+\frac{\partial^2 u}{\partial z^2}=-f(r, z),$$
$$0\leqslant r_0<r<R, \quad 0<z<l.$$

Dans ce cas on peut introduire en direction de r un maillage irrégulier arbitraire et en direction de z un maillage à pas constant h_2.

Si en cas de l'équation de Poisson il s'agit de trouver la solution à la surface du cylindre, c'est-à-dire si

$$\frac{1}{R^2}\frac{\partial^2 u}{\partial \varphi^2}+\frac{\partial^2 u}{\partial z^2}=-f(\varphi, z), \quad 0\leqslant\varphi\leqslant 2\pi, \quad 0<z<l,$$

le problème de différences correspondant se réduit au problème aux limites vectoriel périodique (4), en admettant en direction de z un maillage irrégulier quelconque.

Dans le système de coordonnées polaire, les schémas aux différences admissibles sont les schémas des équations de Poisson dans un cercle, un anneau et dans des secteurs circulaire ou annulaire

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right)+\frac{1}{r^2}\frac{\partial^2 u}{\partial \varphi^2}=-f(r, \varphi), \quad (r, \varphi)\in G.$$

Pour le cercle et l'anneau, le schéma aux différences se réduit au problème périodique (4), tandis que pour les secteurs aux problèmes (1), (2) ou (1), (3). Dans ce cas on peut introduire un maillage irrégulier en direction de r.

A un problème aux limites périodique (4) se réduit également le schéma aux différences de l'équation de Poisson donnée sur la surface d'une sphère de rayon R:

$$\frac{1}{R^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial u}{\partial\theta}\right)+\frac{1}{R^2\sin^2\theta}\frac{\partial^2 u}{\partial\varphi^2}=-f(\varphi, \theta).$$

4. Problème discret de Dirichlet de grand ordre de précision. Voyons maintenant l'exemple d'un schéma aux différences qui se réduit à l'équation vectorielle (5), plus générale que (1). Ecrivons sur un maillage rectangulaire $\bar{\omega}=\{x_{ij}=(ih_1, jh_2)\in\bar{G}, 0\leqslant i\leqslant M, 0\leqslant \leqslant j\leqslant N, h_1M=l_1, h_2N=l_2\}$ le schéma du problème discret de Dirichlet pour l'équation de Poisson de grand ordre de p écision

$$\begin{aligned}&y_{\bar{x}_1x_1}+y_{\bar{x}_2x_2}+\frac{h_1^2+h_2^2}{12}y_{\bar{x}_1x_1\bar{x}_2x_2}=-\varphi(x), \quad x\in\omega,\\ &y(x)=g(x), \quad x\in\gamma.\end{aligned} \tag{28}$$

La solution du schéma aux différences (28) avec le choix adéquat du second membre $\varphi(x)$ converge avec la vitesse $O(h_1^4+h_2^4)$

vers une solution suffisamment lisse du problème différentiel, si $h_1 \neq h_2$, et avec la vitesse $O(h^6)$, si $h_1 = h_2 = h$.

Réduisons (28) au problème aux limites d'une équation vectorielle triponctuelle

$$-BY_{j-1}+AY_j-BY_{j+1}=F_j, \quad 1\leqslant j\leqslant N-1, \qquad Y_0=F_0, \quad Y_N=F_N. \tag{29}$$

Pour cela il faut multiplier (28) par $(-h_2^2)$ et répartir la différence divisée $\left(y+\frac{h_1^2+h_2^2}{12}\,y_{\bar{x}_1x_1}\right)_{\bar{x}_2x_2}$ entre les points en utilisant les notations

$$Y_j = (y\,(1,\ j),\ y\,(2,\ j),\ \ldots,\ y\,(M-1,\ j)),$$

$$F_j = (h_2^2\bar{\varphi}\,(1,\ j),\ h_2^2\varphi\,(2,\ j),\ \ldots,\ h_2^2\varphi\,(M-2,\ j),\ h_2^2\bar{\varphi}\,(M-1,j)),$$

$$1\leqslant j\leqslant N-1,$$

où

$$\bar{\varphi}\,(1,\ j)=\varphi\,(1,\ j)+\frac{1}{h_1^2}\left(g\,(0,\ j)+\frac{h_1^2+h_2^2}{12}\,g_{\bar{x}_2x_2}(0,\ j)\right),$$

$$\bar{\varphi}\,(M-1,\ j)=\varphi\,(M-1,\ j)+\frac{1}{h_1^2}\left(g\,(M,\ j)+\frac{h_1^2+h_2^2}{12}\,g_{\bar{x}_2x_2}(M,\ j)\right)$$

et

$$F_j = (g\,(1,\ j),\ g\,(2,\ j),\ \ldots,\ g\,(M-1,\ j)),\quad j=0,\ N.$$

Dans ce cas les matrices B et A correspondent aux opérateurs de différences Λ_1 et Λ, où

$$\Lambda_1 y=y+\frac{h_1^2+h_2^2}{12}\,y_{\bar{x}_1x_1}, \qquad h_1\leqslant x_1\leqslant l_1-h_1,$$

$$\Lambda y=2y-\frac{5h_2^2-h_1^2}{6}\,y_{\bar{x}_1x_1}, \quad h_1\leqslant x_1\leqslant l_1-h_1,$$

et $y=0$ pour $x_1=0$ et $x_1=l_1$. Ces matrices sont tridiagonales et, comme il est aisé de le vérifier, permutables.

Le problème aux limites (29) peut être réduit au problème (1), (2). A cette fin il faut multiplier chaque équation de (29) à gauche par B^{-1}, si la matrice inverse à B existe. Cherchons la condition suffisante d'existence de B^{-1}. La matrice inverse à B existera apparemment, si le système d'équations algébriques linéaires

$$BY=F \tag{30}$$

possède une solution unique pour tout second membre F.

En vertu de la définition de la matrice B, (30) peut être écrit sous forme de schéma aux différences

$$\Lambda_1 y=y+\frac{h_1^2+h_2^2}{12}\,y_{\bar{x}_1x_1}=f, \quad h_1\leqslant x_1\leqslant l_1-h_1, \qquad y\,(0)=y\,(l_1)=0. \tag{31}$$

Au § 1, ch. II on a montré que si pour le schéma (31) sont remplies les conditions suffisantes de stabilité de la méthode du balayage, la solution de l'équation (31) existe et est unique pour tout second membre f, cette dernière pouvant être obtenue par la méthode du balayage. En répartissant la différence divisée $y_{\bar{x}_1 x_1}$ entre les points, écrivons (31) sous forme des équations scalaires triponctuelles

$$-A_i y_{i-1} + C_i y_i - B_i y_{i+1} = F_i, \quad 1 \leqslant i \leqslant M-1, \qquad y_0 = 0, \quad y_M = 0, \tag{32}$$

où $A_i = B_i = \frac{h_1^2 + h_2^2}{12 h_1^2}$, $C_i = \frac{h_1^2 + h_2^2}{6 h_1^2} - 1$.

Rappelons que pour (32) les conditions suffisantes de stabilité de la méthode du balayage prennent la forme $|C_i| \geqslant |A_i| + |B_i|$, $i = 1, 2, \ldots, M-1$. A partir de ces conditions il s'ensuit que la matrice B possède une matrice inverse au cas où les pas du maillage $\bar{\omega}$ satisfont à la limitation $h_2 \leqslant \sqrt{2} h_1$. Une fois cette condition remplie, le problème (29) peut être réduit au problème (1), (2) avec $C = B^{-1}A$.

§ 2. Méthode de réduction totale pour le premier problème aux limites

1. Procédé d'élimination impair-pair. Passons maintenant à la description de la méthode de réduction totale. Commençons par le premier problème aux limites sur les équations vectorielles triponctuelles

$$-Y_{j-1} + CY_j - Y_{j+1} = F_j, \quad 1 \leqslant j \leqslant N-1, \qquad Y_0 = F_0, \quad Y_N = F_N. \tag{1}$$

L'idée de résolution du problème (1) par la méthode de réduction totale réside dans l'élimination de proche en proche des équations (1) des inconnues Y_j, au départ aux numéros impairs de j, ensuite, des équations restantes aux numéros j multiples de 2, puis de 4, etc. Chaque opération d'élimination abaisse le nombre d'inconnues et si N est la puissance de 2, c'est-à-dire $N = 2^n$, finalement, après élimination, il ne reste qu'une seule équation qui permet de trouver $Y_{N/2}$. La marche par remontée de la méthode consiste dans la recherche de proche en proche des inconnues Y_j, d'abord aux numéros j multiples de $N/4$, puis de $N/8$, $N/16$, etc.

La méthode de réduction totale est apparemment une variante de la méthode d'élimination de Gauss appliquée au problème (1), où l'élimination des inconnues s'effectue dans un ordre déterminé. Rappelons qu'à la différence de cette méthode dans la méthode du balayage matriciel l'élimination des inconnues est réalisée dans l'ordre naturel.

Soit donc $N = 2^n$, $n > 0$. Pour des raisons de simplification. introduisons les notations suivantes: $C^{(0)} = C$, $F_j^{(0)} = F_j$, $j = 1, 2, \ldots, N-1$, qui une fois utilisées permettent d'écrire (1) sous la forme

$$-Y_{j-1} + C^{(0)}Y_j - Y_{j+1} = F_j^{(0)}, \quad 1 \leqslant j \leqslant N-1, N = 2^n,$$
$$Y_0 = F_0, \quad Y_N = F_N. \tag{1'}$$

Etudions la première opération du procédé d'élimination. A ce stade, des équations du système (1'), pour j multiple de 2, éliminons les inconnues Y_j aux numéros impairs de j. Pour cela écrivons trois équations successives de (1'):

$$-Y_{j-2}+C^{(0)}Y_{j-1}-Y_j \quad =F_{j-1}^{(0)},$$
$$-Y_{j-1}+C^{(0)}Y_j \quad -Y_{j+1}=F_j^{(0)},$$
$$-Y_j+C^{(0)}Y_{j+1}-Y_{j+2}=F_{j+1}^{(0)}, \quad j=2, 4, 6, \ldots, N-2.$$

Multiplions la seconde équation à gauche par $C^{(0)}$ et additionnons les trois équations obtenues. Il vient

$$-Y_{j-2}+C^{(1)}Y_j-Y_{j+2}=F_j^{(1)}, \quad j=2, 4, 6, \ldots, N-2,$$
$$Y_0=F_0, \quad Y_N=F_N, \tag{2}$$

où

$$C^{(1)}=[C^{(0)}]^2-2E,$$
$$F_j^{(1)}=F_{j-1}^{(0)}+C^{(0)}F_j^{(0)}+F_{j+1}^{(0)}, \quad j=2, 4, 6, \ldots, N-2.$$

Le système (2) ne contient que des inconnues Y_j aux numéros de j pairs, le nombre d'inconnues dans (2) est de $N/2-1$ et si le système est résolu les inconnues Y_j aux numéros impairs, en vertu de (1'), peuvent être obtenues à partir des équations

$$C^{(0)}Y_j=F_j^{(0)}+Y_{j-1}+Y_{j+1}, \quad j=1, 3, 5, \ldots, N-1 \tag{3}$$

dont les seconds membres sont déjà connus.

Bref, le problème de départ (1') est équivalent au système (2) et aux équations (3), la structure du système (2) étant analogue au système initial.

Au second stade d'élimination des équations du système « raccourci » (2) on élimine pour j multiples de 4 les inconnues aux numéros j multiples de 2, mais non multiples de 4. De façon analogue. au premier stade on prend trois équations du système (2):

$$-Y_{j-4}+C^{(1)}Y_{j-2}-Y_j \quad =F_{j-2}^{(1)},$$
$$-Y_{j-2}+C^{(1)}Y_j \quad -Y_{j+2}=F_j^{(1)},$$
$$-Y_j+C^{(1)}Y_{j+2}-Y_{j+4}=F_{j+2}^{(1)}, \quad j=4, 8, 12, \ldots, N-4.$$

la seconde équation est multipliée à gauche par $C^{(1)}$ et les trois équations s'additionnent. On obtient finalement un système de $N/4-1$ équations contenant les inconnues Y_j aux numéros de j multiples de 4:

$$-Y_{j-4}+C^{(2)}Y_j-Y_{j+4}=F_j^{(2)}, \quad j=4,\ 8,\ 12,\ \ldots,\ N-4,$$
$$Y_0=F_0, \quad Y_N=F_N;$$

les équations $C^{(1)}Y_j = F_j^{(1)} + Y_{j-2} + Y_{j+2}$, $j = 2,\ 6,\ 10,\ \ldots$ $\ldots,\ N-2$ permettant de trouver les inconnues aux numéros multiples de 2 mais non multiples de 4, et les équations (3) fournissant les inconnues aux numéros impairs. Dans ce cas la matrice $C^{(2)}$ et les seconds membres $F_j^{(2)}$ s'obtiennent à l'aide des formules

$$C^{(2)}=[C^{(1)}]^2-2E,$$
$$F_j^{(2)}=F_{j-2}^{(1)}+C^{(1)}F_j^{(1)}+F_{j+2}^{(1)}, \quad j=4,\ 8,\ 12,\ \ldots,\ N-4.$$

Ce procédé d'élimination peut être poursuivi. Au bout de la l-ième opération on obtient le système réduit pour des inconnues aux numéros multiples de 2^l:

$$-Y_{j-2^l}+C^{(l)}Y_j-Y_{j+2^l}=F_j^{(l)}, \quad j=2^l,\ 2\cdot 2^l,\ 3\cdot 2^l,\ \ldots,\ N-2^l,$$
$$Y_0=F_0, \quad Y_N=F_N, \tag{4}$$

et un groupe d'équations

$$C^{(k-1)}Y_j=F_j^{(k-1)}+Y_{j-2^{k-1}}+Y_{j+2^{k-1}},$$
$$j=2^{k-1},\ 3\cdot 2^{k-1},\ 5\cdot 2^{k-1},\ \ldots,\ N-2^{k-1}, \tag{5}$$

qui, une fois résolus successivement pour $k = l,\ l-1,\ \ldots,\ 1$, permettent d'obtenir les inconnues restantes. Les matrices $C^{(k)}$ et les seconds membres $F_j^{(k)}$ s'obtiennent par les formules de récurrence

$$C^{(k)}=[C^{(k-1)}]^2-2E,$$
$$F_j^{(k)}=F_{j-2^{k-1}}^{(k-1)}+C^{(k-1)}F_j^{(k-1)}+F_{j+2^{k-1}}^{(k-1)}, \tag{6}$$
$$j=2^k,\ 2\cdot 2^k,\ 3\cdot 2^k,\ \ldots,\ N-2^k,$$

pour $k = 1,\ 2,\ \ldots$

Il s'ensuit de (4) qu'après la $(n-1)$-ième opération d'élimination $(l = n-1)$ il reste une équation en $Y_{2^{n-1}} = Y_{N/2}$:

$$C^{(n-1)}Y_j=F_j^{(n-1)}+Y_{j-2^{n-1}}+Y_{j+2^{n-1}}=F_j^{(n-1)}+Y_0+Y_N, \quad j=2^{n-1},$$
$$Y_0=F_0, \quad Y_N=F_N$$

avec le second membre connu. En joignant cette équation à (5), on aboutit à ce que toutes les inconnues s'obtiennent de proche en pro-

che à partir des équations

$$C^{(k-1)}Y_j = F_j^{(k-1)} + Y_{j-2^{k-1}} + Y_{j+2^{k-1}}, \quad Y_0 = F_0, \quad Y_N = F_N,$$
$$j = 2^{k-1},\ 3\cdot 2^{k-1},\ 5\cdot 2^{k-1},\ \ldots,\ N-2^{k-1}, \quad k = n,\ n-1,\ \ldots,\ 1. \tag{7}$$

En résumé, les formules (6) et (7) décrivent complètement la méthode de réduction totale. Suivant les formules (6) se transforment les seconds membres, tandis que les équations (7) permettent d'obtenir la solution du problème initial (1).

La méthode décrite sera appelée méthode de réduction totale, car dans ce cas la diminution successive du nombre des équations dans le système est effectuée jusqu'au bout, tant qu'il ne reste qu'une équation pour $Y_{N/2}$. Dans la méthode de réduction partielle, qui sera étudiée au ch. IV, on n'effectue qu'un abaissement partiel de l'ordre du système et le système « raccourci » se résout par une méthode spéciale.

2. Transformation du second membre et inversion des matrices. Le calcul du second membre $F_j^{(k)}$ selon les formules de récurrence (6) peut entraîner l'accumulation d'erreurs de calcul au cas où la norme de la matrice $C^{(k-1)}$ est supérieure à l'unité. En outre, les matrices $C^{(k)}$ sont, à proprement parler, des matrices pleines, même si la matrice de départ $C^{(0)} = C$ était tridiagonale. Or cette circonstance stimule fortement l'accroissement du volume des opérations de calcul de $F_j^{(k)}$ suivant les formules (6). Pour les exemples étudiés au § 1 la norme de la matrice est effectivement beaucoup supérieure à l'unité et un tel algorithme de la méthode se caractérisera par une instabilité des calculs.

Pour obvier à cette difficulté, au lieu des vecteurs $F_j^{(k)}$ calculons les vecteurs $p_j^{(k)}$ qui sont reliés à $F_j^{(k)}$ par la relation suivante:

$$F_j^{(k)} \equiv \prod_{l=0}^{k-1} C^{(l)} p_j^{(k)} 2^k, \tag{8}$$

où l'on posera formellement que $\prod_{l=0}^{-1} C^{(l)} = E$, de sorte que $p_j^{(0)} \equiv \equiv F_j^{(0)} \equiv F_j$.

Cherchons les relations de récurrence qui satisfont à $p_j^{(k)}$. Pour cela portons (8) dans (6). En posant que $C^{(l)}$ est une matrice non dégénérée pour tout l, à partir de (6) il vient

$$2\prod_{l=0}^{k-1} C^{(l)} p_j^{(k)} = \prod_{l=0}^{k-2} C^{(l)} \left[p_{j-2^{k-1}}^{(k-1)} + C^{(k-1)} p_j^{(k-1)} + p_{j+2^{k-1}}^{(k-1)}\right]$$

ou

$$2C^{(k-1)} p_j^{(k)} = p_{j-2^{k-1}}^{(k-1)} + C^{(k-1)} p_j^{(k-1)} + p_{j+2^{k-1}}^{(k-1)}. \tag{9}$$

En posant $S_j^{(k-1)} = 2p_j^{(k)} - p_j^{(k-1)}$, de (9) on tire que $p_j^{(k)}$ peuvent être obtenus successivement suivant les formules suivantes:

$$C^{(k-1)}S_j^{(k-1)} = p_{j-2^{k-1}}^{(k-1)} + p_{j+2^{k-1}}^{(k-1)}, \quad p_j^{(k)} = 0{,}5\,(p_j^{(k-1)} + S_j^{(k-1)}),$$
$$j = 2^k,\ 2\cdot 2^k,\ 3\cdot 2^k,\ \ldots,\ N-2^k, \quad k = 1,\ 2,\ \ldots,\ n-1, \tag{10}$$
$$p_j^{(0)} \equiv F_j.$$

Les relations de récurrence (10) contiennent une addition des vecteurs, une multiplication d'un vecteur par un nombre et une inversion de la matrice $C^{(k-1)}$.

Il ne reste qu'à éliminer $F_j^{(k-1)}$ des équations (7). En portant (8) dans (7), on obtient

$$C^{(k-1)}Y_j = 2^{k-1}\prod_{l=0}^{k-2} C^{(l)}p_j^{(k-1)} + Y_{j-2^{k-1}} + Y_{j+2^{k-1}},$$
$$Y_0 = F_0, \quad Y_N = F_N, \tag{11}$$
$$j = 2^{k-1},\ 3\cdot 2^{k-1},\ \ldots,\ N-2^{k-1}, \quad k = n,\ n-1,\ \ldots,\ 1.$$

Il est également nécessaire ici d'inverser les matrices $C^{(k-1)}$, mais en outre dans le second membre de (11) est apparue une multiplication de la matrice par un vecteur. Dans l'algorithme étudié plus bas le procédé utilisé d'inversion de la matrice $C^{(k-1)}$ permet de se débarrasser de la fâcheuse opération de multiplication par un vecteur et de réduire la mise en œuvre de (11) à une inversion de matrices et une addition de vecteurs.

Voyons maintenant le problème de l'inversion des matrices $C^{(k-1)}$ déterminées au moyen des formules de récurrence (6)

$$C^{(k)} = [C^{(k-1)}]^2 - 2E, \quad k = 1,\ 2,\ \ldots,\ C^{(0)} = C. \tag{12}$$

Il s'ensuit de (12) que $C^{(k)}$ est un polynôme matriciel de puissance 2^k en C avec coefficient unitaire près de la puissance majeure. Au moyen des polynômes connus de Tchébychev ce polynôme s'exprime de la façon suivante:

$$C^{(k)} = 2T_{2^k}\left(\frac{1}{2}C\right), \quad k = 0,\ 1,\ \ldots, \tag{13}$$

où $T_n(x)$ est le polynôme de Tchébychev de n-ième degré de première espèce (voir point 2, § 4, ch. 1):

$$T_n(x) = \begin{cases} \cos(n \arccos x), & |x| \leqslant 1, \\ \frac{1}{2}\left[(x+\sqrt{x^2-1})^n + (x+\sqrt{x^2-1})^{-n}\right], & |x| \geqslant 1. \end{cases}$$

En effet, en vertu des propriétés du polynôme $T_n(x)$

$$T_{2n}(x) = 2[T_n(x)]^2 - 1, \quad T_1(x) = x,$$

(13) découle de façon évidente de (12).

Ensuite, en utilisant la relation

$$\prod_{l=0}^{k-2} 2T_{2^l}(x) = U_{2^{k-1}-1}(x),$$

liant les polynômes de Tchébychev de première espèce aux polynômes de deuxième espèce $U_n(x)$, où

$$U_n(x) = \begin{cases} \dfrac{\sin((n+1)\arccos x)}{\sin(\arccos x)}, & |x| \leqslant 1, \\ \dfrac{1}{2\sqrt{x^2-1}}[(x+\sqrt{x^2-1})^{n+1} - (x+\sqrt{x^2-1})^{-(n+1)}], & |x| \geqslant 1, \end{cases}$$

on calcule sans peine le produit des polynômes $C^{(l)}$

$$\prod_{l=0}^{k-2} C^{(l)} = U_{2^{k-1}-1}\left(\frac{1}{2}C\right). \tag{14}$$

Bref, les expressions explicites de $C^{(k)}$ et $\prod_{l=0}^{k-1} C^{(l)}$ sont obtenues.

Dans la suite il nous faut utiliser le lemme 6 (voir point 5, § 4, ch. II). Selon le lemme 6, tout rapport de polynômes $g_m(x)/f_n(x)$ sans racines communes, au cas de $n > m$, et de racines simples $f_n(x)$ se décompose en fractions élémentaires de la façon suivante:

$$\frac{g_m(x)}{f_n(x)} = \sum_{l=1}^{n} \frac{a_l}{x-x_l}, \quad a_l = \frac{g_m(x_l)}{f'_n(x_l)},$$

où x_l sont les racines du polynôme $f_n(x)$.

Utilisons le lemme 6 pour la décomposition des rapports $1/T_n(x)$ et $U_{n-1}(x)/T_n(x)$ en fractions simples. Les racines du polynôme $T_n(x)$ sont connues:

$$x_l = \cos\frac{(2l-1)}{2n}\pi, \quad l = 1, 2, \ldots, n, \tag{15}$$

et en ces points le polynôme $U_{n-1}(x)$ prend des valeurs non nulles

$$U_{n-1}(x_l) = \frac{\sin(n\arccos x_l)}{\sin(\arccos x_l)} = \frac{(-1)^{l+1}}{\sin\frac{(2l-1)}{2n}\pi}, \quad l = 1, 2, \ldots, n.$$

Aussi en utilisant la relation $T'_n(x) = nU_{n-1}(x)$ du lemme 6 obtient-on le développement suivant:

$$\frac{1}{T_n(x)} = \sum_{l=1}^{n} \frac{(-1)^{l+1}\sin\frac{(2l-1)\pi}{2n}}{n(x-x_l)}, \tag{16}$$

$$\frac{U_{n-1}(x)}{T_n(x)} = \sum_{l=1}^{n} \frac{1}{n(x-x_l)}, \tag{17}$$

où x_l est défini dans (15). Les développements cherchés sont trouvés.

Cherchons maintenant les expressions des matrices $[C^{(k-1)}]^{-1}$ et $[C^{(k-1)}]^{-1}\prod\limits_{l=0}^{k-2} C^{(l)}$ au moyen de la matrice C. A partir de (13) et (14), compte tenu des développements des polynômes algébriques (16), (17), il vient

$$[C^{(k-1)}]^{-1}=\sum_{l=1}^{2^{k-1}}\alpha_{l,\,k-1}\left(C-2\cos\frac{(2l-1)\pi}{2^k}E\right)^{-1},$$

$$[C^{(k-1)}]^{-1}\prod_{l=0}^{2^{k-1}}C^{(l)}=\frac{1}{2^{k-1}}\sum_{l=1}^{2^{k-1}}\left(C-2\cos\frac{(2l-1)\pi}{2^k}E\right)^{-1}.$$

Les relations obtenues permettent d'écrire sous la forme suivante les formules (10) :

$$\begin{aligned}&S_j^{(k-1)}=\prod_{l=1}^{2^{k-1}}\alpha_{l,\,k-1}C_{l,\,k-1}^{-1}\,(p_{j-2^{k-1}}^{(k-1)}+p_{j+2^{k-1}}^{(k-1)}),\\&p_j^{(k)}=0{,}5\,(p_j^{(k-1)}+S_j^{(k-1)}),\\&p_j^{(0)}\equiv F_j,\\&j=2^k,\ 2\cdot 2^k,\ 3\cdot 2^k,\ \ldots,\ N-2^k,\quad k=1,\ 2,\ \ldots,\ n-1,\end{aligned}\qquad(18)$$

ainsi que les formules (11) :

$$\begin{aligned}&Y_j=\prod_{l=1}^{2^{k-1}}C_{l,\,k-1}^{-1}\,[p_j^{(k-1)}+\alpha_{l,\,k-1}(Y_{j-2^{k-1}}+Y_{j+2^{k-1}})],\\&Y_0=F_0,\quad Y_N=F_N,\\&j=2^{k-1},\ 3\cdot 2^{k-1},\ 5\cdot 2^{k-1},\ \ldots,\ N-2^{k-1},\\&k=n,\ n-1,\ \ldots,\ 1,\end{aligned}\qquad(19)$$

avec les notations

$$C_{l,\,k-1}=C-2\cos\frac{(2l-1)\pi}{2^k}E,\quad \alpha_{l,\,k-1}=\frac{(-1)^{l+1}}{2^{k-1}}\sin\frac{(2l-1)\pi}{2^k}.\qquad(20)$$

En résumé, on a obtenu les formules transformées (18), (19) décrivant la méthode de résolution de (1) par réduction totale. Ces formules ne contiennent que des opérations d'addition des vecteurs, de multiplication d'un vecteur par un nombre et d'inversion des matrices.

Notons que si C est une matrice tridiagonale, toute matrice $C_{l,\,k-1}$ sera également une matrice tridiagonale. Le problème de

l'inversion de ces matrices a été résolu au chapitre II. Ensuite, si pour la matrice C est remplie la condition $(CY, Y) \geqslant 2\,(Y, Y)$, il s'ensuit alors de (20) que les matrices $C_{l,\,k}$ seront définies positives et, par suite, auront des inverses limités. On obtient alors du développement $[C^{(k-1)}]^{-1}$ que pour tout $k \geqslant 1$ les matrices $C^{(k-1)}$ ne sont pas dégénérées. Rappelons que cette hypothèse a été utilisée pour obtenir les formules (10).

3. Algorithme de la méthode. Les formules (18), (19) obtenues plus haut servent de base au premier algorithme de la méthode. Voyons tout d'abord quelles grandeurs intermédiaires et à quel moment doivent être calculées et mémorisées à des utilisations ultérieures.

L'analyse des formules (19) montre qu'en fixant k pour le calcul de Y_j, on utilise les vecteurs $p_j^{(k-1)}$ aux numéros de $j = 2^{k-1}$, $3\cdot 2^{k-1}, \ldots, N - 2^{k-1}$. Tout vecteur $p_j^{(l)}$ au même numéro j mais à numéro l inférieur à $k - 1$ est un vecteur auxiliaire et se mémorise

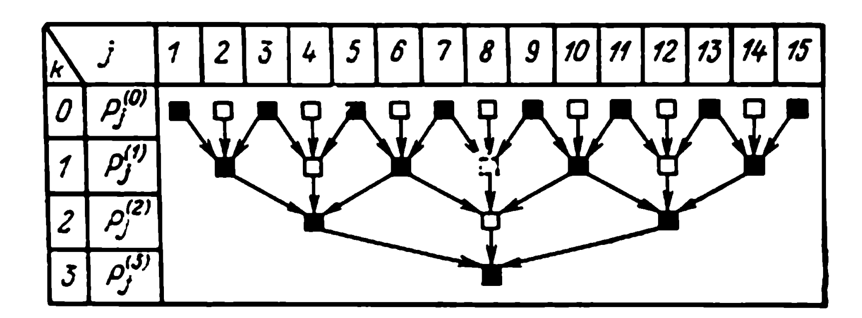


Fig. 1.

temporairement. Aussi les vecteurs $p_j^{(k)}$, déterminés à la k-ième opération suivant (18), peuvent occuper la place de $p_j^{(k-1)}$, de même que les inconnues Y_j calculées avec (19). La méthode n'oblige pas de prévoir une mémoire supplémentaire pour l'ordinateur : tous les vecteurs $p_j^{(k)}$ se placent à l'endroit où se placera par la suite Y_j.

Illustrons sur un exemple l'organisation des calculs dans l'algorithme étudié. Soit $N = 16$ $(n = 4)$. Sur la figure 1 on a indiqué la succession des calculs et de mémorisation des vecteurs $p_j^{(k)}$. Le carré en noir signifie que pour la valeur indiquée de l'indice k on mémorise pour l'utilisation ultérieure le vecteur $p_j^{(k)}$ au numéro j correspondant. Et, respectivement, le carré blanc signifie que $p_j^{(k)}$ est un vecteur auxiliaire et n'est mémorisé à l'endroit indiqué qu'à titre temporaire. On indique par des flèches les vecteurs $p_i^{(k-1)}$ utilisés pour le calcul de $p_j^{(k)}$.

Dans le sens direct, la méthode permet de mémoriser les vecteurs $p_j^{(k)}$ suivants:

$$p_1^{(0)},\ p_2^{(1)},\ p_3^{(0)},\ p_4^{(2)},\ p_5^{(0)},\ p_6^{(1)},\ p_7^{(0)},\ p_8^{(3)},\ p_9^{(0)},$$
$$p_{10}^{(1)},\ p_{11}^{(0)},\ p_{12}^{(2)},\ p_{13}^{(0)},\ p_{14}^{(1)},\ p_{15}^{(0)}.$$

Ils sont utilisés pour le calcul de Y_j, la méthode étant appliquée par remontée.

La figure 2 montre l'ordre suivi pour le calcul des inconnues Y_j (représentées symboliquement par ○). Les flèches indiquent les

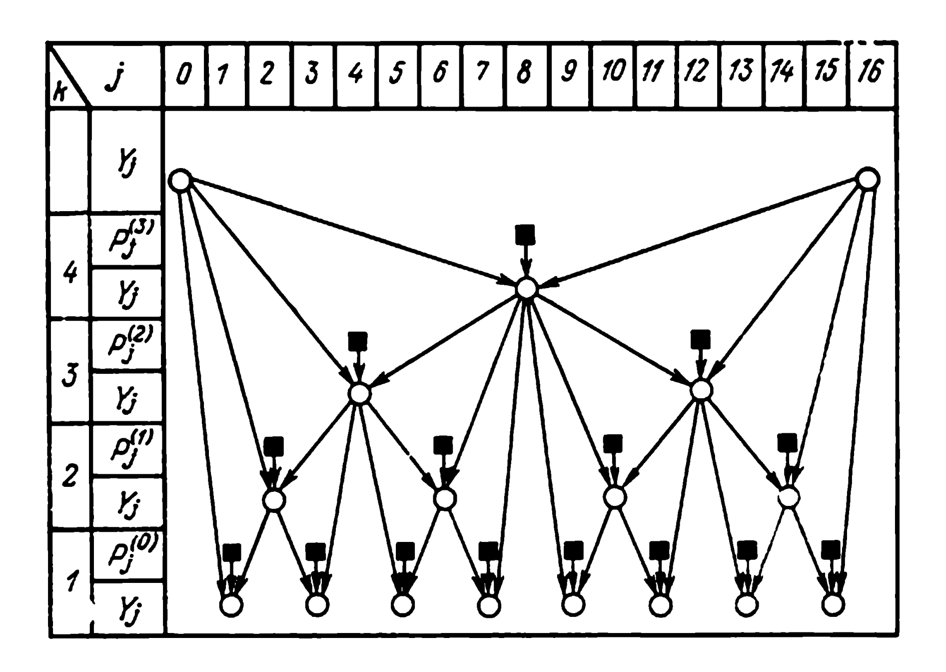


Fig. 2.

Y_j obtenus lors des opérations précédentes et les $p_j^{(k-1)}$ (représentés symboliquement par ■) utilisés pour le calcul de Y_j, k étant donné.

Passons maintenant à la description de l'algorithme de la méthode de réduction totale. Dans le sens direct, selon (18), la méthode est mise en œuvre de la façon suivante:

1) On fixe les valeurs pour $p_j^{(0)} = F_j$, $j = 1, 2, \ldots, N-1$.

2) Pour chaque $k = 1, 2, \ldots, n-1$ fixé pour un $j = 2^k, 2\cdot 2^k, \ldots, N-2^k$ donné, on calcule et l'on mémorise d'abord les vecteurs

$$\varphi = p_{j-2^{k-1}}^{(k-1)} + p_{j+2^{k-1}}^{(k-1)}. \tag{21}$$

Ensuite, pour $l = 1, 2, \ldots, 2^{k-1}$ on résout les équations

$$C_{l,\,k-1} v_l = \alpha_{l,\,k-1}\varphi. \tag{22}$$

Finalement, par accumulation graduelle des résultats à l'endroit de $p_j^{(k-1)}$, on obtient $p_j^{(k)}$

$$p_j^{(k)} = 0{,}5(p_j^{(k-1)} + v_1 + v_2 + \ldots + v_{2^{k-1}}). \qquad (23)$$

Par remontée, selon (19), la méthode est mise en œuvre de la façon suivante:

1) On fixe les valeurs de Y_0 et Y_N : $Y_0 = F_0$, $Y_N = F_N$.

2) Pour chaque $k = n, n-1, \ldots, 1$ fixé pour un $j = 2^{k-1}$, $3\cdot 2^{k-1}$, $5\cdot 2^{k-1}$, $\ldots$, $N - 2^{k-1}$ donné, on calcule et l'on mémorise les vecteurs

$$\varphi = Y_{j-2^{k-1}} + Y_{j+2^{k-1}}, \quad \psi = p_j^{(k-1)}. \qquad (24)$$

Ensuite, pour $l = 1, 2, \ldots, 2^{k-1}$, on résout les équations

$$C_{l,\,k-1} v_l = \psi + \alpha_{l,\,k-1}\varphi. \qquad (25)$$

Finalement, par l'accumulation graduelle des valeurs à l'endroit de $p_j^{(k-1)}$, on trouve le vecteur des inconnues Y_j

$$Y_j = v_1 + v_2 + \ldots + v_{2^{k-1}}. \qquad (26)$$

Passons maintenant au calcul du nombre d'opérations arithmétiques que vaut la mise en œuvre de l'algorithme décrit. Soit M la dimension du vecteur d'inconnues Y_j, $\mathring{q}$ désignant le nombre d'opérations exigées pour la résolution de l'équation de la forme (22) ou (25), le second membre étant donné. Admettons que les quantités $\alpha_{l,\,k}$ ont été déjà trouvées.

Calculons maintenant le nombre d'opérations arithmétiques Q_1 que coûte la méthode en sens direct. Pour des k et j fixés, le calcul du vecteur φ suivant les formules (21) exige M opérations. Ensuite, pour chaque l, le calcul du second membre dans (22) et la résolution de l'équation (22) il faut $M + \mathring{q}$ opérations. Aussi pour l'obtention de tous les v_l faut-il $2^{k-1}(M + \mathring{q})$ opérations. Le calcul de $p_j^{(k)}$ suivant la formule (23) coûte $2^{k-1} M + M$ opérations. En résumé, pour le calcul de $p_j^{(k)}$ pour un k et un j donnés il faut $M + 2^{k-1}(2M + \mathring{q})$ opérations.

Ensuite, pour chaque k fixé il faut calculer $p_j^{(k)}$ pour $N/2^k - 1$ cas différents. Donc le nombre total d'opérations Q_1 que coûte la mise en œuvre de la méthode en sens direct vaut

$$Q_1 = \sum_{k=1}^{n-1} [M + (2M + \mathring{q})\, 2^{k-1}] \left(\frac{N}{2^k} - 1\right) =$$
$$= (M + 0{,}5\mathring{q})\, Nn - (M + \mathring{q})\, N - M\,(n-1) + \mathring{q}. \qquad (27)$$

Calculons maintenant le nombre d'opérations Q_2 que coûte la mise en œuvre de la méthode par remontée. Pour des k et j fixés le calcul suivant les formules (24) exige M opérations, l'obtention de tous les v_l avec (25) $(2M+\overset{\bullet}{q})2^{k-1}$ opérations et le calcul de Y_j suivant la formule (26) $(2^{k-1}-1)M$ opérations. Comme le nombre de différentes valeurs de j pour lesquelles, pour un k fixé, les opérations mentionnées sont effectuées est $N/2^k$, Q_2 vaut

$$Q_2=\sum_{k=1}^{n}[M+(2M+\overset{\circ}{q})\,2^{k-1}+(2^{k-1}-1)\,M]\,\frac{N}{2^k}=$$

$$=(1{,}5M+0{,}5\overset{\circ}{q})\,Nn. \qquad (28)$$

En additionnant (27) et (28), compte tenu de ce que $n=\log_2 N$, on obtient l'estimation suivante du nombre d'opérations que coûte la méthode de réduction totale mise en œuvre suivant l'algorithme décrit plus haut

$$Q=Q_1+Q_2=(2{,}5M+\overset{\bullet}{q})\,N\log_2 N-(M+\overset{\bullet}{q})\,N-$$

$$-M\,(n-1)+\overset{\bullet}{q}. \qquad (29)$$

Il s'ensuit de (29) que si $\overset{\bullet}{q}=O\,(M)$, $Q=O\,(MN\log_2 N)$.

4. Second algorithme de la méthode. L'avantage principal de l'algorithme construit réside dans ses exigences minimales envers la mémoire de l'ordinateur : il n'exige pas de mémoire supplémentaire pour la mémorisation de l'information auxiliaire. Cet avantage s'acquiert au prix d'une certaine augmentation du volume des calculs réitérés des grandeurs intermédiaires. Examinons encore un algorithme de la méthode se caractérisant par un moindre volume des calculs mais exigeant une mémoire supplémentaire comparable en puissance au nombre total d'inconnues dans le problème.

Pour construire le second algorithme, reprenons les formules (6), (7) décrivant la méthode de réduction totale :

$$C^{(k)}=[C^{(k-1)}]^2-2E,$$

$$F_j^{(k)}=F_{j-2^{k-1}}^{(k-1)}+C^{(k-1)}F_j^{(k-1)}+F_{j+2^{k-1}}^{(k-1)}, \qquad (6')$$

$$j=2^k,\ 2\cdot 2^k,\ 3\cdot 2^k,\ \ldots,\ N-2^k,\quad k=1,\ 2,\ \ldots,\ n-1,$$

$$C^{(k-1)}Y_j=F_j^{(k-1)}+Y_{j-2^{k-1}}+Y_{j+2^{k-1}},$$

$$Y_0=F_0,\quad Y_N=F_N, \qquad (7')$$

$$j=2^{k-1},\ 3\cdot 2^{k-1},\ 5\cdot 2^{k-1},\ \ldots,\ N-2^{k-1},\qquad k=n,\ n-1,\ \ldots,\ 1.$$

Ici, comme au cas du premier algorithme, les vecteurs $F_j^{(k)}$ ne sont pas calculés directement et à leur place on détermine les vecteurs

$p_j^{(k)}$ et $q_j^{(k)}$ reliés à $F_j^{(k)}$ par la relation suivante:

$$F_j^{(k)} = C^{(k)} p_j^{(k)} + q_j^{(k)},$$
$$j = 2^k,\ 2\cdot 2^k,\ 3\cdot 2^k,\ \ldots,\ N-2^k, \quad k = 0,\ 1,\ \ldots,\ n-1. \tag{30}$$

Cherchons les formules de récurrence servant au calcul des vecteurs $p_j^{(k)}$ et $q_j^{(k)}$. Vu qu'au lieu du vecteur $F_j^{(k)}$ on a introduit deux vecteurs, il y a un certain arbitraire dans la détermination de $p_j^{(k)}$ et $q_j^{(k)}$. Choisissons $p_j^{(0)}$ et $q_j^{(0)}$ de manière à satisfaire à la condition initiale $F_j^{(0)} \equiv F_j$. Pour ce faire, posons

$$p_j^{(0)} = 0, \quad q_j^{(0)} = F_j, \quad j = 1,\ 2,\ \ldots,\ N-1. \tag{31}$$

Ensuite, en portant (30) dans (6′), on obtient

$$C^{(k)} p_j^{(k)} + q_j^{(k)} = C^{(k-1)} [q_j^{(k-1)} + p_{j-2^{k-1}}^{(k-1)} +$$
$$+ C^{(k-1)} p_j^{(k-1)} + p_{j+2^{k-1}}^{(k-1)}] + q_{j-2^{k-1}}^{(k-1)} + q_{j+2^{k-1}}^{(k-1)},$$
$$j = 2^k,\ 2\cdot 2^k,\ \ldots,\ N-2^k, \quad k = 1,\ 2,\ \ldots,\ n-1.$$

En choisissant

$$q_j^{(k)} = 2p_j^{(k)} + q_{j-2^{k-1}}^{(k-1)} + q_{j+2^{k-1}}^{(k-1)} \tag{32}$$

et compte tenu de ce que $C^{(k)} + 2E = [C^{(k-1)}]^2$, on en tire

$$C^{(k-1)} p_j^{(k)} = q_j^{(k-1)} + p_{j-2^{k-1}}^{(k-1)} + C^{(k-1)} p_j^{(k-1)} + p_{j+2^{k-1}}^{(k-1)}. \tag{33}$$

Ici on admet de nouveau que $C^{(l)}$ est une matrice non dégénérée pour tout l.

En posant $S_j^{(k-1)} = p_j^{(k)} - p_j^{(k-1)}$, on obtient de (31)-(33) les formules de récurrence suivantes qui permettent de calculer les vecteurs $p_j^{(k)}$ et $q_j^{(k)}$:

$$\begin{gathered} C^{(k-1)} S_j^{(k-1)} = q_j^{(k-1)} + p_{j-2^{k-1}}^{(k-1)} + p_{j+2^{k-1}}^{(k-1)}, \\ p_j^{(k)} = p_j^{(k-1)} + S_j^{(k-1)}, \\ q_j^{(k)} = 2p_j^{(k)} + q_{j-2^{k-1}}^{(k-1)} + q_{j+2^{k-1}}^{(k-1)}, \\ q_j^{(0)} \equiv F_j, \quad p_j^{(0)} \equiv 0, \end{gathered} \tag{34}$$
$$j = 2^k,\ 2\cdot 2^k,\ 3\cdot 2^k,\ \ldots,\ N-2^k, \quad k = 1,\ 2,\ \ldots,\ n-1.$$

Il ne reste qu'à éliminer $F_j^{(k-1)}$ de la formule (7′). Portant (30) dans (7′) et posant $t_j^{(k-1)} = Y_j - p_j^{(k-1)}$, on obtient les formules suivantes pour le calcul de Y_j:

$$\begin{gathered} C^{(k-1)} t_j^{(k-1)} = q_j^{(k-1)} + Y_{j-2^{k-1}} + Y_{j+2^{k-1}}, \\ Y_j = p_j^{(k-1)} + t_j^{(k-1)}, \\ Y_0 = F_0, \quad Y_N = F_N, \end{gathered} \tag{35}$$
$$j = 2^{k-1},\ 3\cdot 2^{k-1},\ 5\cdot 2^{k-1},\ \ldots,\ N-2^{k-1}, \quad k = n,\ n-1,\ \ldots,\ 1.$$

Bref, on a obtenu les formules (34), (35) sur lesquelles s'appuie le second algorithme de la méthode de réduction totale. Ces formules comprennent des opérations d'addition des vecteurs et d'inversion des matrices $C^{(k-1)}$.

Arrêtons-nous maintenant sur le problème de l'inversion des matrices $C^{(k-1)}$. Comme il a été montré plus haut, la matrice $C^{(k)}$ est un polynôme de degré 2^k relativement à la matrice initiale C et se détermine à l'aide de la formule (13) au moyen du polynôme de Tchébychev de première espèce $T_n(x)$:

$$C^{(k)} = 2T_{2^k}\left(\frac{1}{2}C\right),$$

le coefficient près de la puissance majeure étant l'unité. Vu que les racines du polynôme $T_n(x)$ sont connues (voir (15)), on peut représenter $C^{(k)}$ sous forme factorisée

$$C^{(k)} = \prod_{l=1}^{2^k}\left(C - 2\cos\frac{(2l-1)\pi}{2^{k+1}}E\right), \quad k = 0, 1, \ldots$$

En utilisant les notations (20), on peut écrire la matrice $C^{(k-1)}$ sous la forme suivante:

$$C^{(k-1)} = \prod_{l=1}^{2^{k-1}} C_{l,k-1}, \quad C_{l,k-1} = C - 2\cos\frac{(2l-1)\pi}{2^k}E. \tag{36}$$

La factorisation (36) permet de résoudre sans peine les équations de la forme $C^{(k-1)}v = \varphi$ avec le second membre fixé φ. L'algorithme suivant permet de résoudre ce problème par inversion des facteurs dans (36):

$$v_0 = \varphi, \quad C_{l,k-1}v_l = v_{l-1}, \quad l = 1, 2, \ldots, 2^{k-1},$$

avec $v = v_{2^{k-1}}$. Cet algorithme sera utilisé pour l'inversion des matrices $C^{(k-1)}$.

Décrivons maintenant le second algorithme de la méthode de réduction totale. Dans le sens direct la méthode est mise en œuvre sur la base de (34) de la façon suivante:

1) On fixe les valeurs pour $q_j^{(0)}$: $q_j^{(0)} = F_j$, $j = 1, 2, \ldots, N-1$.

2) La première opération pour $k = 1$ est réalisée séparément suivant les formules tenant compte des données initiales $p_j^{(0)} \equiv 0$. On résout les équations pour $p_j^{(1)}$ et on calcule $q_j^{(1)}$:

$$\begin{aligned} &Cp_j^{(1)} = q_j^{(0)}, \\ &q_j^{(1)} = 2p_j^{(1)} + q_{j-1}^{(0)} + q_{j+1}^{(0)}, \quad j = 2, 4, 6, \ldots, N-2. \end{aligned} \tag{37}$$

3) Pour chaque $k = 2, 3, \ldots, n - 1$ fixé on calcule en mémorisant les vecteurs

$$v_j^{(0)} = q_j^{(k-1)} + p_{j-2^{k-1}}^{(k-1)} + p_{j+2^{k-1}}^{(k-1)}, \quad j = 2^k, 2\cdot 2^k, 3\cdot 2^k, \ldots, N-2^k. \tag{38}$$

Ensuite, avec $l = 1, 2, 3, \ldots, 2^{k-1}$ fixé pour chaque $j = 2^k$, $2\cdot 2^k$, $3\cdot 2^k$, $\ldots$, $N - 2^k$ on résout les équations

$$C_{l,k-1} v_j^{(l)} = v_j^{(l-1)} \tag{39}$$

possédant la même matrice, mais avec des seconds membres différents. On obtient ainsi les vecteurs $v_j^{(2^{k-1})}$ (dans les formules (34) à ces vecteurs correspondent $S_j^{(k-1)}$). Les vecteurs $p_j^{(k)}$ et $q_j^{(k)}$ sont calculés à l'aide des formules

$$\begin{gathered} p_j^{(k)} = p_j^{(k-1)} + v_j^{(2^{k-1})}, \\ q_j^{(k)} = 2p_j^{(k)} + q_{j-2^{k-1}}^{(k-1)} + q_{j+2^{k-1}}^{(k-1)}, \\ j = 2^k, 2\cdot 2^k, 3\cdot 2^k, \ldots, N-2^k. \end{gathered} \tag{40}$$

La méthode est mise en œuvre par remontée suivant (35):

1) On fixe les valeurs pour Y_0 et Y_N: $Y_0 = F_0$, $Y_N = F_N$.

2) Pour chaque $k = n, n - 1, \ldots, 2$ fixé on calcule en mémorisant les vecteurs

$$\begin{gathered} v_j^{(0)} = q_j^{(k-1)} + Y_{j-2^{k-1}} + Y_{j+2^{k-1}}, \\ j = 2^{k-1}, 3\cdot 2^{k-1}, 5\cdot 2^{k-1}, \ldots, N-2^{k-1}. \end{gathered} \tag{41}$$

Ensuite, avec $l = 1, 2, \ldots, 2^{k-1}$ fixé, pour chaque $j = 2^{k-1}$, $3\cdot 2^{k-1}$, $5\cdot 2^{k-1}$, $\ldots$, $N - 2^{k-1}$, on résout les équations

$$C_{l,k-1} v_j^{(l)} = v_j^{(l-1)}. \tag{42}$$

Finalement on trouve les vecteurs $v_j^{(2^{k-1})}$ (dans (35) il leur correspond les vecteurs $t_j^{(k-1)}$). Ensuite, on calcule Y_j suivant la formule

$$Y_j = p_j^{(k-1)} + v_j^{(2^{k-1})}, \quad j = 2^{k-1}, 3\cdot 2^{k-1}, 5\cdot 2^{k-1}, \ldots, N-2^{k-1}. \tag{43}$$

3) L'opération finale de la méthode par remontée consiste, pour $k = 1$, à rechercher la solution de l'équation

$$CY_j = q_j^{(0)} + Y_{j-1} + Y_{j+1}, \qquad j = 1, 3, 5, \ldots, N-1. \tag{44}$$

R e m a r q u e s u r l' a l g o r i t h m e. Tous les vecteurs $p_j^{(k)}$ redéfinis à l'aide des formules (37) et (40) se disposent à la place des $p_j^{(k-1)}$. Tous les vecteurs $v_j^{(l)}$ des formules (38), (39), (41), (42), les vecteurs $q_j^{(k)}$ redéfinis suivant les formules (37), (40), de même que la solution Y_j s'ensuivant de (43) et (44), se placent à l'endroit

de $q_j^{(k-1)}$. Donc cet algorithme exige que la mémoire de l'ordinateur soit 1,5 fois plus puissante que le nombre d'inconnues dans le problème.

L'abaissement du volume des calculs dans l'algorithme étudié par rapport à celui exigé par le premier algorithme est dû au fait qu'avec la résolution de la série de problèmes (39) et (42) pour des j différents avec les mêmes matrices $C_{l,\,k-1}$ tous les calculs ne sont effectués que pour le premier problème de la série, la résolution de chaque problème suivant comportant un nombre sensiblement inférieur d'opérations arithmétiques. Donnons le nombre d'opérations que coûte le second algorithme en désignant par $\mathring{q}$ le nombre d'opérations exigées pour la résolution de l'équation de la forme (39) ou (42) avec le second membre fixé, et par $\bar{q}$ le nombre d'opérations exigées pour la résolution de la même équation, mais avec un autre second membre $(\bar{q} < \mathring{q})$.

Le nombre d'opérations effectuées pour la mise en œuvre de la méthode dans le sens direct vaut

$$Q_1 = \sum_{k=1}^{n-1} \left\{6M\left(\frac{N}{2^k} - 1\right) + \left[\mathring{q} + \bar{q}\left(\frac{N}{2^k} - 2\right)\right]2^{k-1}\right\} - \\ - 3M\left(\frac{N}{2} - 1\right) = 0{,}5\bar{q}Nn + (0{,}5\mathring{q} - 1{,}5\bar{q} + 4{,}5M)\,N - \\ - 6Mn - (\mathring{q} - 2\bar{q} + 3M),$$

et par remontée

$$Q_2 = \sum_{k=1}^{n} \left\{3M\frac{N}{2^k} + \left[\mathring{q} + \left(\frac{N}{2^k} - 1\right)\bar{q}\right]2^{k-1}\right\} - \frac{MN}{2} = \\ = 0{,}5\bar{q}Nn + (\mathring{q} - \bar{q} + 2{,}5M)\,N - \mathring{q} + \bar{q} - 3M.$$

Le nombre total d'opérations que coûte le second algorithme vaut

$$Q = Q_1 + Q_2 = \\ = \bar{q}N\log_2 N + (1{,}5\mathring{q} - 2{,}5\bar{q} + 7M)\,N - 6Mn - 2\mathring{q} + 3\bar{q} - 6M. \quad (45)$$

Il s'ensuit de l'estimation (45) que si $\mathring{q} = O(M)$, $\bar{q} = O(M)$ et $Q = O(MN\log_2 N)$, le coefficient près du terme principal $MN\log_2 N$ étant ici plus petit que dans l'estimation (29), car $\bar{q} < \mathring{q}$.

Arrêtons-nous brièvement sur une autre particularité du second algorithme. Si dans le premier algorithme l'inversion des matrices $C^{(k-1)}$ s'effectuait par inversion des facteurs $C_{l,\,k-1}$ et par l'addition subséquente des résultats, dans le second algorithme on procède à l'inversion des facteurs de proche en proche et l'on obtient le ré-

sultat après inversion du dernier facteur. Sous l'angle du processus réel des calculs, qui tient compte des erreurs d'arrondi, l'ordre d'inversion des facteurs $C_{l,\,k-1}$ dans le second algorithme est essentiel. On se heurtera à une situation analogue au chapitre VI avec l'étude de la méthode itérative de Tchébychev.

On peut recommander l'ordre suivant d'inversion des matrices $C_{l,\,k-1}$. Faisons correspondre à la matrice $C^{(k-1)}$ le vecteur $\theta_{2^{k-1}}$ de dimension 2^{k-1} et dont les composantes sont des nombres entiers allant de 1 à 2^{k-1}. Soit

$$\theta_{2^{k-1}} = \{\theta_{2^{k-1}}(1),\ \theta_{2^{k-1}}(2),\ \ldots,\ \theta_{2^{k-1}}(2^{k-1})\},$$

autrement dit, le l-ième élément du vecteur $\theta_{2^{k-1}}$ est désigné par $\theta_{2^{k-1}}(l)$. Le nombre $\theta_{2^{k-1}}(l)$ définit l'ordre d'inversion de la matrice $C_{l,\,k-1}$.

Le vecteur $\theta_{2^{k-1}}$ se construit par récurrence. Soit $\theta_2 = \{2,\ 1\}$. Alors le procédé de doublement de la dimension du vecteur se décrit de la façon suivante:

$$\theta_{2m} = \{\theta_{2m}(4i-3) = \theta_m(2i-1),\quad \theta_{2m}(4i-2) = \theta_m(2i-1) + m,$$
$$\theta_{2m}(4i-1) = \theta_m(2i) + m,\quad \theta_{2m}(4i) = \theta_m(2i),$$
$$i = 1,\ 2,\ \ldots,\ m/2\},\quad m = 2,\ 4,\ 8,\ \ldots$$

Exemple: $\theta_{16} = \{2,\ 10,\ 14,\ 6,\ 8,\ 16,\ 12,\ 4,\ 3,\ 11,\ 15,\ 7,\ 5,\ 13,\ 9,\ 1\}$ et, par conséquent, la matrice $C_{6,\,16}$ sera invertie lors de la seizième inversion, tandis que la matrice $C_{12,\,16}$ lors de la septième.

§ 3. Exemples d'application de la méthode

1. Problème discret de Dirichlet pour l'équation de Poisson dans un rectangle. Examinons comment s'applique la méthode de réduction totale élaborée plus haut à la résolution du problème discret de Dirichlet pour l'équation de Poisson dans un rectangle. Comme il a été montré auparavant, le problème de différences

$$y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} = -\varphi(x),\qquad x \in \omega,$$
$$y(x) = g(x),\qquad x \in \gamma,$$

donné sur le maillage rectangulaire $\bar{\omega} = \{x_{ij} = (ih_1,\ jh_2),\ 0 \leqslant i \leqslant M,\ 0 \leqslant j \leqslant N,\ h_1 M = l_1,\ h_2 N = l_2\}$, s'écrit sous forme du premier problème aux limites pour des équations vectorielles triponctuelles

$$\begin{aligned} &-Y_{j-1} + CY_j - Y_{j+1} = F_j,\quad 1 \leqslant j \leqslant N-1,\\ &Y_0 = F_0,\quad Y_N = F_N. \end{aligned}\tag{1}$$

$$Y_j = (y(1,\ j),\quad y(2,\ j),\ \ldots,\ y(M-1,\ j)),\qquad 0 \leqslant j \leqslant N,$$

est ici le vecteur des inconnues dont les composantes sont les valeurs de la fonction de maille $y(i, j)$ sur la j-ième ligne du maillage,

$$F_j = (h_2^2\bar{\varphi}(1, j), h_2^2\varphi(2, j), \ldots, h_2^2\varphi(M-2, j), h_2^2\bar{\varphi}(M-1, j)),$$
$$1 \leqslant j \leqslant N-1,$$

$$F_j = (g(1, j), g(2, j), \ldots, g(M-1, j)), \qquad j = 0, N,$$

où

$$\bar{\varphi}(1, j) = \varphi(1, j) + \frac{1}{h_1^2} g(0, j),$$

$$\bar{\varphi}(M-1, j) = \varphi(M-1, j) + \frac{1}{h_1^2} g(M, j).$$

La matrice carrée C correspond à l'opérateur de différences Λ, où

$$\Lambda y = 2y - h_2^2 y_{\bar{x}_1 x_1}, \quad h_1 \leqslant x_1 \leqslant l_1 - h_1,$$
$$y = 0, \quad x_1 = 0, l_1,$$

de sorte que

$$CY_j = (\Lambda y(1, j), \Lambda y(2, j), \ldots, \Lambda y(M-1, j)).$$

Le problème (1) se prête à la résolution par l'un des deux algorithmes de la méthode de réduction totale. La phase principale de ces algorithmes est la résolution des équations du type

$$C_{l, k-1} V = F, \quad C_{l, k-1} = C - 2\cos\frac{(2l-1)\pi}{2^k} E \tag{2}$$

avec le second membre fixé F. V est ici le vecteur des inconnues, $V = (v(1), v(2), \ldots, v(M-1))$ de dimension $M-1$ (pour simplifier l'écriture, l'indice de V et de F a été négligé).

Rappelons que le nombre d'opérations que coûte la résolution du problème (1) suivant le premier algorithme est fonction du nombre d'opérations $\overset{\circ}{q}$ exigé par la résolution de l'équation (2) (voir (29) point 3, § 2), et suivant le second algorithme, du nombre d'opérations supplémentaires $\bar{q}$ exigé par la résolution de l'équation (2), mais munie d'un autre second membre (voir (45) point 4, § 2).

Donnons pour l'exemple considéré le procédé de résolution de l'équation (2) et fournissons l'estimation pour $\overset{\circ}{q}$ et $\bar{q}$. Il s'ensuit de la définition de la matrice C que la résolution de l'équation (2) équivaut à la recherche de la solution du problème de différences suivant :

$$\begin{aligned} &2\left(1 - \cos\frac{(2l-1)\pi}{2^k}\right) v - h_2^2 v_{\bar{x}_1 x_1} = f(i), \quad 1 \leqslant i \leqslant M-1, \\ &v(0) = v(M) = 0, \end{aligned} \tag{3}$$

où $f(i) = f_i$ est l'i-ième composante du vecteur F. En répartissant la différence divisée $v_{\bar{x}_1 x_1}$ entre les points, écrivons (3) sous forme d'équation aux différences triponctuelle ordinaire pour des inconnues scalaires $v(i) = v_i$:

$$-v_{i-1} + av_i - v_{i+1} = bf_i, \quad 1 \leqslant i \leqslant M - 1, \qquad v_0 = v_M = 0, \tag{4}$$

où $a = 2\left[1 + b\left(1 - \cos\frac{(2l-1)\pi}{2^k}\right)\right]$, $b = \frac{h_1^2}{h_2^2}$. Le problème (4) est un cas spécial de problèmes aux limites triponctuels, dont les méthodes de résolution ont été étudiées au chapitre II. On a montré que la méthode efficace de résolution des problèmes de la forme (4) est la méthode du balayage. Donnons les formules de calculs de la méthode du balayage appliquées au problème (4):

$$\alpha_{i+1} = 1/(a - \alpha_i), \qquad i = 1, 2, \ldots, M - 1, \qquad \alpha_1 = 0,$$
$$\beta_{i+1} = (bf_i + \beta_i)\,\alpha_{i+1}, \qquad i = 1, 2, \ldots, M - 1, \qquad \beta_1 = 0,$$
$$v_i = \alpha_{i+1} v_{i+1} + \beta_{i+1}, \qquad i = M - 1, M - 2, \ldots, 1, \quad v_M = 0.$$

Il s'ensuit de ces formules que le problème (4) et, partant, l'équation (2) peuvent être résolus pour a et b donnés en $\dot{q} = 7(M - 1)$ opérations. La résolution de l'équation (2) avec un autre second membre F n'implique pas un recalcul des coefficients de balayage α_i et, par suite, le nombre d'opérations supplémentaires $\bar{q}$ vaut $\bar{q} = 5(M - 1)$. Ces opérations serviront au calcul de β_i et à la recherche de la solution v_i. Notons que la méthode du balayage sera numériquement stable pour (4), de sorte que la condition suffisante de stabilité de la méthode envers les erreurs d'arrondi, ayant la forme de $a \geqslant 2$, est remplie.

En portant $\dot{q}$ dans l'estimation (29) fournie au point 3, § 2 du nombre d'opérations du premier algorithme, on obtient, en retenant les principaux termes, $Q^{(1)} \approx 9{,}5\, MN \log_2 N - 8MN$. Pour le second algorithme sur la base de l'estimation (45) du point 4, § 2, on obtient l'estimation suivante du nombre d'opérations: $Q^{(2)} \approx \approx 5MN \log_2 N + 5MN$. En résumé, pour chacun des algorithmes considérés le nombre d'opérations de la méthode de réduction totale, appliquée à la résolution du problème discret de Dirichlet sur l'équation de Poisson dans un rectangle, est une quantité de l'ordre de $O(MN \log_2 N)$, le second algorithme comportant moins d'opérations arithmétiques. Par exemple, pour $M = N = 64$, on obtient $Q^{(1)} \approx 1{,}4Q^{(2)}$ et pour $M = N = 128$ respectivement $Q^{(1)} \approx \approx 1{,}46Q^{(2)}$.

On ne donnera pas les formules de calculs pour l'algorithme de résolution du problème de différences mentionné, car au niveau vectoriel elles ont été décrites en détail au § 2.

Au point 2 du § 1 on a fourni des exemples d'autres problèmes de différences aux limites qui se ramènent au problème (1). Ils diffèrent du problème de Dirichlet étudié par le type de conditions aux limites sur les côtés du rectangle pour $x_1 = 0$ et $x_1 = l_1$, ce qui conduit à des matrices C différentes. C'est ainsi que pour le problème (10)-(12) du point 2, § 1, aux conditions aux limites de troisième ou de deuxième espèce pour $x_1 = 0,\ l_1$, l'équation (2) est équivalente au problème de différences

$$2\left(1-\cos\frac{(2l-1)\pi}{2^k}\right)v-h_2^2 v_{\bar{x}_1x_1}=f, \qquad 1\leqslant i\leqslant M-1,$$

$$2\left(1+\frac{h_2^2}{h_1}\varkappa_{-1}-\cos\frac{(2l-1)\pi}{2^k}\right)v-\frac{2h_2^2}{h_1}v_{x_1}=f, \qquad i=0,$$

$$2\left(1+\frac{h_2^2}{h_1}\varkappa_{+1}-\cos\frac{(2l-1)\pi}{2^k}\right)v+\frac{2h_2^2}{h_1}v_{\bar{x}_1}=f, \qquad i=M.$$

Ce problème prend dans la forme triponctuelle ordinaire l'aspect

$$\begin{aligned}&-v_{i-1}+av_i-v_{i+1}=bf_i, \quad 1\leqslant i\leqslant M-1,\\ &v_0=\bar{\varkappa}_1v_1+\mu_1,\\ &v_M=\bar{\varkappa}_2v_{M-1}+\mu_2,\end{aligned} \tag{5}$$

où

$$\bar{\varkappa}_1=\frac{2}{a+2h_1\varkappa_{-1}}, \quad \bar{\varkappa}_2=\frac{2}{a+2h_1\varkappa_{+1}}, \quad \mu_1=\frac{bf_0}{a+2h_1\varkappa_{-1}},$$
$$\mu_2=\frac{bf_M}{a+2h_1\varkappa_{+1}},$$

a et b étant définis plus haut.

Vu que $a>2$ et $\varkappa_{\pm1}\geqslant 0$, on a $0<\bar{\varkappa}_1<1$ et $0<\bar{\varkappa}_2<1$ et la méthode du balayage servant à la résolution du problème (5) sera de même stable, quant à l'algorithme de la méthode de réduction totale, il exigera dans ce cas $O\,(MN\log_2 N)$ opérations arithmétiques.

2. Problème discret de Dirichlet de grand ordre de précision. Au point 4, § 1 le problème discret de Dirichlet pour l'équation de Poisson de grand ordre de précision

$$y_{\bar{x}_1x_1}+y_{\bar{x}_2x_2}+\frac{h_1^2+h_2^2}{12}y_{\bar{x}_1x_1\bar{x}_2x_2}=-\varphi(x), \quad x\in\omega,$$
$$y(x)=g(x), \quad x\in\gamma$$

a été ramené au premier problème aux limites pour une équation vectorielle triponctuelle non réduite

$$\begin{aligned}&-BY_{j-1}+AY_j-BY_{j+1}=F_j, \quad 1\leqslant j\leqslant N-1,\\ &Y_0=F_0, \quad Y_N=F_N.\end{aligned} \tag{6}$$

Les matrices carrées B et A de dimension $(M-1)\times(M-1)$ correspondent aux opérateurs de différences Λ_1 et Λ, où

$$\Lambda_1 y = y + \frac{h_1^2+h_2^2}{12} y_{\bar{x}_1 x_1}, \quad h_1 \leqslant x_1 \leqslant l_1 - h_1,$$
$$\Lambda y = 2y - \frac{5h_2^2-h_1^2}{6} y_{\bar{x}_1 x_1}, \quad h_1 \leqslant x_1 \leqslant l_1 - h_1$$

et $y=0$ pour $x_1=0$ et $x_1=l_1$.

On a montré que si la condition $h_2 \leqslant \sqrt{2}h_1$ est remplie, les équations (6) se réduisent à la forme standard

$$-Y_{j-1} + CY_j - Y_{j+1} = \Phi_j, \quad 1 \leqslant j \leqslant N-1, \qquad Y_0 = \Phi_0, \quad Y_N = \Phi_N, \tag{7}$$

où $C=B^{-1}A$, $\Phi_j = B^{-1}F_j$, $1 \leqslant j \leqslant N-1$ et $\Phi_j = F_j$ pour $j = 0, N$. En outre, on a noté que les matrices A et B sont permutables.

Pour résoudre (7), utilisons le premier algorithme de la méthode. Comme la matrice $C_{l,k-1}$ peut être transcrite sous la forme

$$C_{l,\,k-1} = C - 2\cos\frac{(2l-1)\pi}{2^k} E = B^{-1}\left(A - 2\cos\frac{(2l-1)\pi}{2^k} B\right),$$

les formules (18), (19) du § 2, définissant le premier algorithme, prennent l'aspect suivant:

$$S_j^{(k-1)} = \sum_{l=1}^{2^{k-1}} \alpha_{l,\,k-1}\left(A - 2\cos\frac{(2l-1)\pi}{2^k} B\right)^{-1} B\,(p_{j-2^{k-1}}^{(k-1)} + p_{j+2^{k-1}}^{(k-1)}),$$
$$p_j^{(k)} = 0{,}5\,(p_j^{(k-1)} + S_j^{(k-1)}),$$
$$j = 2^k,\ 2\cdot 2^k, \ldots, N-2^k, \quad k = 1, 2, \ldots, n-1,$$
$$Bp_j^{(0)} \equiv F_j,$$
$$Y_j = \sum_{l=1}^{2^{k-1}} \left(A - 2\cos\frac{(2l-1)\pi}{2^k} B\right)^{-1} B\,[p_j^{(k-1)} + \alpha_{l,\,k-1}(Y_{j-2^{k-1}} + Y_{j+2^{k-1}})],$$
$$Y_0 = F_0, \quad Y_N = F_N, \quad j = 2^{k-1},\ 3\cdot 2^{k-1}, \ldots, N-2^{k-1},$$
$$k = n,\ n-1, \ldots, 1.$$

Pour échapper à l'inversion de la matrice B avec la fixation de $p_j^{(0)}$ et à la multiplication de $p_j^{(k-1)}$ par la matrice B lors du calcul de Y_j, effectuons les substitutions en posant $\bar{p}_j^{(k)} = Bp_j^{(k)}$, $\bar{S}_j^{(k)} = BS_j^{(k)}$. Alors compte tenu de la permutabilité des matrices A et B et, partant, des matrices $\left(1 - 2\cos\frac{(2l-1)\pi}{2^k} B\right)^{-1}$

et B, les formules écrites précédemment prendront la forme (le trait au-dessus de $\bar{p}_j^{(k)}$ et de $\bar{S}_j^{(k)}$ est négligé) :

$$S_j^{(k-1)} = \sum_{l=1}^{2^{k-1}} \alpha_{l,\,k-1} \left(A - 2\cos\frac{(2l-1)\pi}{2^k} B\right)^{-1} B\,(p_{j-2^{k-1}}^{(k-1)} + p_{j+2^{k-1}}^{(k-1)}),$$

$$p_j^{(k)} = 0{,}5\,(p_j^{(k-1)} + S_j^{(k-1)}), \quad j = 2^k,\ 2\cdot 2^k, \ldots, N - 2^k,$$
$$k = 1, 2, \ldots, n-1,$$

$$p_j^{(0)} \equiv F_j,$$

$$Y_j = \sum_{l=1}^{2^{k-1}} \left(A - 2\cos\frac{(2l-1)\pi}{2^k} B\right)^{-1} [p_j^{(k-1)} + \\ + \alpha_{l,\,k-1} B\,(Y_{j-2^{k-1}} + Y_{j+2^{k-1}})],$$

$$Y_0 = F_0, \quad Y_N = F_N, \quad j = 2^{k-1},\ 3\cdot 2^{k-1}, \ldots, N - 2^{k-1},$$
$$k = n,\ n-1, \ldots, 1.$$

Les formules obtenues entraînent les modifications suivantes dans le premier algorithme: la formule (21) du § 2 est remplacée par

$$\varphi = B\,(p_{j-2^{k-1}}^{(k-1)} + p_{j+2^{k-1}}^{(k-1)}),$$

et, au lieu des équations (22), on résout les équations

$$\left(A - 2\cos\frac{(2l-1)\pi}{2^k} B\right) v_l = \alpha_{l,\,k-1}\varphi$$

avec φ calculé. De façon analogue (24) est remplacé par

$$\varphi = B\,(Y_{j-2^{k-1}} + Y_{j+2^{k-1}}), \quad \psi = p_j^{(k-1)},$$

et, au lieu de (25), on résout les équations

$$\left(A - 2\cos\frac{(2l-1)\pi}{2^k} B\right) v_l = \psi + \alpha_{l,\,k-1}\varphi.$$

La phase principale de l'algorithme est donc pour le problème considéré la résolution des équations de la forme

$$\left(A - 2\cos\frac{(2l-1)\pi}{2^k} B\right) V = F \tag{8}$$

avec le second membre F fixé. En utilisant la définition des matrices A et B à l'aide des opérateurs de différences Λ et Λ_1, on obtient que (8) est équivalent à la solution du problème de différences suivant:

$$2\left(1 - \cos\frac{(2l-1)\pi}{2^k}\right) v - \left(\frac{5h_2^2 - h_1^2}{6} + \frac{h_1^2 + h_2^2}{6}\cos\frac{(2l-1)\pi}{2^k}\right) v_{\bar{x}_1 x_1} = f,$$
$$1 \leqslant i \leqslant M-1, \quad v_0 = v_M = 0. \tag{9}$$

En répartissant cette équation entre les points, on obtient le premier problème aux limites sur l'équation scalaire triponctuelle

$$\begin{aligned} &-v_{i-1} + av_i - v_{i+1} = bf_i, \quad 1 \leqslant i \leqslant M-1, \\ &v_0 = v_M = 0, \end{aligned} \tag{10}$$

où

$$a = 2\left[1 + b\left(1 - \cos\frac{(2l-1)\pi}{2^k}\right)\right],$$

$$b = \frac{6h_1^2}{5h_2^2 - h_1^2 + (h_1^2 + h_2^2)\cos\frac{(2l-1)\pi}{2^k}}.$$

Le problème de différences (10) peut être résolu par la méthode du balayage qui s'est avérée numériquement stable au cas où la condition suffisante $|a| \geqslant 2$ est remplie. Montrons que pour tous h_1 et h_2 cette condition est satisfaite. En effet, si h_1 et h_2 sont tels que l'inégalité

$$\frac{h_2^2}{h_1^2} \geqslant \frac{1 - \cos\frac{(2l-1)\pi}{2^k}}{5 + \cos\frac{(2l-1)\pi}{2^k}}, \tag{11}$$

est remplie, alors $0 < b \leqslant \infty$ et, partant, $a > 2$. Remarquons qu'en cas d'égalité dans (11) le coefficient près de $v_{\bar{x}_1 x_1}$ dans (9) devient nul et v peut être obtenu de (9) suivant la formule explicite.

Si (11) n'est pas satisfaite, pour b se vérifie l'estimation

$$b < -6 \Big/ \left(1 - \cos\frac{(2l-1)\pi}{2^k}\right),$$

et, par suite, $a < -10$. La proposition est démontrée.

Bref, pour la résolution du problème discret de Dirichlet de grand ordre de précision on peut recourir à la méthode de réduction totale avec l'estimation $O(MN \log_2 N)$ du nombre d'opérations arithmétiques.

§ 4. Méthode de réduction totale appliquée à d'autres problèmes aux limites

1. Second problème aux limites. On a étudié plus haut la méthode de réduction totale appliquée à la résolution du premier problème aux limites pour les équations vectorielles triponctuelles. On abordera l'étude de la méthode appliquée à des problèmes aux limites plus compliqués par l'étude du *second problème aux limites*. Soit qu'il s'agit de trouver la solution du problème suivant:

$$\begin{aligned} &CY_0 - 2Y_1 = F_0, && j = 0, \\ &-Y_{j-1} + CY_j - Y_{j+1} = F_j, && 1 \leqslant j \leqslant N-1, \\ &-2Y_{N-1} + CY_N = F_N, && j = N, \end{aligned} \tag{1}$$

où $N = 2^n$, $n > 0$.

Le procédé d'élimination successive des inconnues dans (1) est mis en œuvre de la même façon qu'au cas de conditions aux limites de première espèce. A savoir, pour des j pairs on aura les équations

$$-Y_{j-2}+C^{(1)}Y_j-Y_{j+2}=F_j^{(1)},\quad j=2,4,6,\ldots,N-2, \tag{2}$$

et pour des j impairs les équations

$$C^{(0)}Y_j=F_j^{(0)}+Y_{j-1}+Y_{j+1},\quad j=1,3,5,\ldots,N-1, \tag{3}$$

où, comme précédemment, sont utilisées les notations

$$F_j^{(1)}=F_{j-1}^{(0)}+C^{(0)}F_j^{(0)}+F_{j+1}^{(0)},\quad C^{(1)}=[C^{(0)}]^2-2E,$$
$$C^{(0)}=C,\quad F_j^{(0)}\equiv F_j.$$

Sont restées non transformées seules les équations du système (1) pour $j=0$ et $j=N$. Eliminons de ces équations les inconnues Y_j aux numéros j impairs. A cette fin utilisons deux équations voisines. Ecrivons les équations pour $j=0$ et $j=1$:

$$C^{(0)}Y_0-2Y_1=F_0^{(0)},\quad -Y_0+C^{(0)}Y_1-Y_2=F_1^{(0)}.$$

Multiplions la première équation à gauche par $C^{(0)}$ et la seconde par 2, additionnons les équations obtenues et il vient

$$C^{(1)}Y_0-2Y_2=F_0^{(1)}, \tag{4}$$

où $F_0^{(1)}=C^{(0)}F_0^{(0)}+2F_1^{(0)}$. De façon analogue on obtient l'équation

$$-2Y_{N-2}+C^{(1)}Y_N=F_N^{(1)}, \tag{5}$$

où $F_N^{(1)}=2F_{N-1}^{(0)}+C^{(0)}F_N^{(0)}$.

En réunissant (2), (4) et (5), on obtient le système « raccourci » complet d'équations pour des inconnues aux numéros j pairs, ayant une structure analogue à (1):

$$C^{(1)}Y_0-2Y_2=F_0^{(1)},\qquad j=0,$$
$$-Y_{j-2}+C^{(1)}Y_j-Y_{j+2}=F_j^{(1)},\qquad j=2,4,6,\ldots N-2,$$
$$-Y_{N-2}+C^{(1)}Y_N=F_N^{(1)},\qquad j=N,$$

et un groupe d'équations (3) pour des inconnues aux numéros j impairs.

En poursuivant le procédé décrit d'élimination des inconnues, après la n-ième opération d'élimination, on obtient le système pour Y_0 et Y_N:

$$C^{(n)}Y_0-2Y_N=F_0^{(n)},\quad -2Y_0+C^{(n)}Y_N=F_N^{(n)} \tag{6}$$

et des équations permettant de déterminer les inconnues restantes :

$$C^{(k-1)}Y_j = F_j^{(k-1)} + Y_{j-2^{k-1}} + Y_{j+2^{k-1}}. \tag{7}$$

$$j = 2^{k-1},\ 3\cdot 2^{k-1},\ 5\cdot 2^{k-1},\ \ldots,\ N-2^{k-1},\quad k = n,\ n-1,\ \ldots,\ 1,$$

où $F_j^{(k)}$ et $C^{(k)}$ se déterminent par récurrence pour $k = 1, 2, \ldots, n$:

$$\begin{aligned} &F_0^{(k)} = C^{(k-1)}F_0^{(k-1)} + 2F_{2^{k-1}}^{(k-1)},\\ &F_j^{(k)} = F_{j-2^{k-1}}^{(k-1)} + C^{(k-1)}Y_j + F_{j+2^{k-1}}^{(k-1)},\\ &j = 2^k,\ 2\cdot 2^k,\ 3\cdot 2^k,\ \ldots,\ N-2^k,\\ &F_N^{(k)} = 2F_{N-2^{k-1}}^{(k-1)} + C^{(k-1)}F_N^{(k-1)},\\ &C^{(k)} = [C^{(k-1)}]^2 - 2E. \end{aligned} \tag{8}$$

Bref, il faut résoudre le système (6) et, ensuite, à partir des équations (7), obtenir toutes les inconnues restantes.

Ici comme au cas du second algorithme de la méthode de réduction totale, utilisée pour le premier problème aux limites, au lieu des vecteurs $F_j^{(k)}$ on déterminera les vecteurs $p_j^{(k)}$ et $q_j^{(k)}$ associés à $F_j^{(k)}$ par la relation

$$F_j^{(k)} = C^{(k)}p_j^{(k)} + q_j^{(k)}, \tag{9}$$

$$j = 0,\ 2^k,\ 2\cdot 2^k,\ 3\cdot 2^k,\ \ldots,\ N-2^k,\ N,\quad k = 0,\ 1,\ \ldots,\ n.$$

A partir de (8) on obtient, comme auparavant, que $p_j^{(k)}$ et $q_j^{(k)}$, pour $j \neq 0, N$, peuvent être trouvés suivant les formules

$$\begin{aligned} &C^{(k-1)}S_j^{(k-1)} = q_j^{(k-1)} + p_{j-2^{k-1}}^{(k-1)} + p_{j+2^{k-1}}^{(k-1)},\\ &p_j^{(k)} = p_j^{(k-1)} + S_j^{(k-1)},\\ &q_j^{(k)} = 2p_j^{(k)} + q_{j-2^{k-1}}^{(k-1)} + q_{j+2^{k-1}}^{(k-1)},\\ &j = 2^k,\ 2\cdot 2^k,\ \ldots,\ N-2^k,\quad k = 1, 2, \ldots, n-1,\\ &q_j^{(0)} \equiv F_j,\quad p_j^{(0)} \equiv 0. \end{aligned} \tag{10}$$

Cherchons maintenant les formules pour $p_j^{(k)}$ et $q_j^{(k)}$ avec $j = 0, N$. Portant (9) avec $j = 0$ dans (8) pour $F_0^{(k)}$, il vient

$$C^{(k)}p_0^{(k)} + q_0^{(k)} = C^{(k-1)}[q_0^{(k-1)} + 2p_{2^{k-1}}^{(k-1)} + C^{(k-1)}p_0^{(k-1)}] + 2q_{2^{k-1}}^{(k-1)}.$$

En choisissant $q_0^{(k)} = 2p_0^{(k)} + 2q_{2^{k-1}}^{(k-1)}$ et compte tenu de l'égalité (12) du point 1, § 2, on obtient l'équation pour $p_0^{(k)}$

$$C^{(k-1)}p_0^{(k)} = C^{(k-1)}p_0^{(k-1)} + q_0^{(k-1)} + 2p_{2^{k-1}}^{(k-1)}.$$

Bref, les vecteurs $p_0^{(k)}$ et $q_0^{(k)}$ peuvent être obtenus à l'aide des formules de récurrence suivantes:

$$\begin{aligned} &C^{(k-1)} S_0^{(k-1)} = q_0^{(k-1)} + 2p_{2^{k-1}}^{(k-1)}, \\ &p_0^{(k)} = p_0^{(k-1)} + S_0^{(k-1)}, \\ &q_0^{(k)} = 2p_0^{(k)} + 2q_{2^{k-1}}^{(k-1)}, \quad k = 1, 2, \ldots, n, \\ &q_0^{(0)} = F_0, \quad p_0^{(0)} = 0. \end{aligned} \tag{11}$$

Les formules pour $p_N^{(k)}$ et $q_N^{(k)}$ s'obtiennent de façon analogue:

$$\begin{aligned} &C^{(k-1)} S_N^{(k-1)} = q_N^{(k-1)} + 2p_{N-2^{k-1}}^{(k-1)}, \\ &p_N^{(k)} = p_N^{(k-1)} + S_N^{(k-1)}, \\ &q_N^{(k)} = 2p_N^{(k)} + 2q_{N-2^{k-1}}^{(k-1)}, \quad k = 1, 2, \ldots, n, \\ &q_N^{(0)} = F_N, \quad p_N^{(0)} = 0. \end{aligned} \tag{12}$$

Les formules (10)-(12) permettent donc d'obtenir complètement tous les vecteurs $p_j^{(k)}$ et $q_j^{(k)}$ cherchés. Il ne reste qu'à éliminer $F_j^{(k)}$ de (6) et (7). Portant (9) dans (7), on obtient les formules suivantes pour le calcul de Y_j:

$$\begin{aligned} &C^{(k-1)} t_j^{(k-1)} = q_j^{(k-1)} + Y_{j-2^{k-1}} + Y_{j+2^{k-1}}, \\ &Y_j = p_j^{(k-1)} + t_j^{(k-1)}, \end{aligned} \tag{13}$$

$$j = 2^{k-1}, \quad 3 \cdot 2^{k-1}, \ 5 \cdot 2^{k-1}, \ \ldots, \ N - 2^{k-1},$$

$$k = n, \ n-1, \ \ldots, \ 1.$$

Il ne reste qu'à trouver Y_0 et Y_N à partir de (6). Toutefois, notons d'abord que de (11) et (12) pour $k = n$ s'ensuivent les égalités

$$q_0^{(n)} = 2p_0^{(n)} + 2q_{2^{n-1}}^{(n-1)}, \quad q_N^{(n)} = 2p_N^{(n)} + 2q_{2^{n-1}}^{(n-1)},$$

c'est-à-dire

$$q_0^{(n)} - q_N^{(n)} = 2\,(p_0^{(n)} - p_N^{(n)}). \tag{14}$$

Ensuite, de (9) et (14) on obtient que

$$F_0^{(n)} - F_N^{(n)} = C^{(n)}\,(p_0^{(n)} - p_N^{(n)}) + q_0^{(n)} - q_N^{(n)} =$$

$$= (C^{(n)} + 2E)\,(p_0^{(n)} - p_N^{(n)}).$$

En tenant compte de la formule (12) du point 1, § 2 on aura finalement:

$$F_0^{(n)} - F_N^{(n)} = [C^{(n-1)}]^2\,(p_0^{(n)} - p_N^{(n)}). \tag{15}$$

Profitons de la relation obtenue pour trouver Y_0 et Y_N à partir de (6). En soustrayant de la première équation du système (6) la seconde équation, compte tenu de (15) et de l'égalité (12) du point 1, § 2, on obtient que

$$(C^{(n)}+2E)(Y_0-Y_N)=[C^{(n-1)}]^2(Y_0-Y_N)=F_0^{(n)}-F_N^{(n)}=$$
$$=[C^{(n-1)}]^2(p_0^{(n)}-p_N^{(n)}).$$

En admettant que $C^{(n-1)}$ est une matrice non dégénérée, il vient de l'expression précédente

$$Y_0=Y_N+p_0^{(n)}-p_N^{(n)}. \qquad (16)$$

En portant Y_0 trouvé dans la seconde équation du système (9), on obtient l'équation permettant de trouver Y_N:

$$B^{(n)}Y_N=F_N^{(n)}+2(p_0^{(n)}-p_N^{(n)})=B^{(n)}p_N^{(n)}+q_N^{(n)}+2p_0^{(n)},$$

où $B^{(n)}=C^{(n)}-2E$. Par conséquent, si l'on pose $t^{(n)}=Y_N-p_N^{(n)}$, on est en mesure de trouver Y_N, en résolvant l'équation

$$B^{(n)}t^{(n)}=q_N^{(n)}+2p_0^{(n)}, \quad (Y_N=p_N^{(n)}+t^{(n)}). \qquad (17)$$

A partir de (16) on obtient que Y_0 peut être trouvé suivant la formule

$$Y_0=p_0^{(n)}+t^{(n)}, \qquad (18)$$

où $t^{(n)}$ est connu.

En résumé, les formules (10)-(12), (17) et (18) décrivent la méthode de réduction totale permettant de résoudre le second problème aux limites sur les équations vectorielles triponctuelles (1).

R e m a r q u e. Si Y_0 est donné, c'est-à-dire si au lieu du problème (1) on résout le problème

$$-Y_{j-1}+CY_j-Y_{j+1}=F_j, \quad 1\leqslant j\leqslant N-1,$$
$$-2Y_{N-1}+CY_N=F_N, \quad j=N, \quad Y_0=F_0,$$

il n'est pas nécessaire de calculer les vecteurs $p_0^{(k)}$ et $q_0^{(k)}$, et Y_N, comme cela s'ensuit de (6) et (9), s'obtient en résolvant l'équation

$$C^{(n)}t_N^{(n)}=q_N^{(n)}+2Y_0, \qquad (Y_N=p_N^{(n)}+t_N^{(n)}).$$

De façon analogue, si Y_N est donné, il n'est pas nécessaire de calculer les vecteurs $p_N^{(k)}$ et $q_N^{(k)}$, tandis que Y_0 se détermine à partir de l'équation $C^{(n)}t_0^{(n)}=q_0^{(n)}+2Y_N$, $Y_0=p_0^{(n)}+t_0^{(n)}$.

Pour terminer la description de la méthode de réduction il faut indiquer les procédés d'inversion des matrices $C^{(k)}$ et $B^{(n)}=C^{(n)}-2E$. Pour l'inversion des matrices $C^{(k-1)}$ on utilise la factorisa-

tion obtenue plus haut (voir (36) § 2)

$$C^{(k-1)} = \prod_{l=1}^{2^{k-1}} C_{l,\,k-1}, \quad C_{l,\,k-1} = C - 2\cos\frac{(2l-1)\pi}{2^k}E. \tag{19}$$

Notons qu'en satisfaisant à la condition $(CY, Y) \geqslant 2(Y, Y)$, toutes les matrices $C_{l,k-1}$ sont non dégénérées et, par suite, la matrice $C^{(k-1)}$ est également non dégénérée. Considérons plus en détail le problème de l'inversion de la matrice $B^{(n)}$.

De la définition de $B^{(n)}$ et de la relation (12) du point 1, § 2 il vient

$$B^{(n)} = C^{(n)} - 2E = [C^{(n-1)}]^2 - 4E = (C^{(n-1)} + 2E)(C^{(n-1)} - 2E) =$$
$$= [C^{(n-2)}]^2 [C^{(n-1)} - 2E] = \ldots = [C^{(n-2)}C^{(n-3)} \ldots C^{(0)}]^2 (C^{(1)} - 2E) =$$
$$= [C^{(n-2)}C^{(n-3)} \ldots C^{(0)}]^2 [C^{(0)} - 2E)(C^{(0)} + 2E) =$$
$$= [\prod_{k=1}^{n-1} C^{(k-1)}]^2 (C - 2E)(C + 2E).$$

En y portant (19), on obtient la représentation suivante de la matrice:

$$B^{(n)} = [\prod_{k=1}^{n-1} \prod_{l=1}^{2^{k-1}} C_{l,\,k-1}]^2 (C - 2E)(C + 2E). \tag{20}$$

La matrice $B^{(n)}$ est ainsi factorisée et l'inversion de $B^{(n)}$ peut être réalisée de proche en proche par inversion des facteurs.

R e m a r q u e 1. On peut aboutir à une écriture plus condensée de (20):

$$B^{(n)} = \prod_{l=1}^{n} \left(C - 2\cos\frac{l\pi}{2^{n-1}}E\right).$$

R e m a r q u e 2. Il s'ensuit de (20) que la matrice $B^{(n)}$ sera non dégénérée si la condition $(CY, Y) > 2(Y, Y)$ est remplie. Si, par contre, il existe un vecteur tel que $Y^* \neq 0$, pour lequel $CY^* = 2Y^*$, $B^{(n)}$ est alors dégénérée et l'application directe de la méthode de réduction totale s'avère impossible. C'est la conséquence de la dégénérescence dans le cas considéré de la matrice du système (1). En effet, dans ce cas le système homogène (1) a une solution non nulle $Y_j = Y^*$ et, par suite, le système (1) n'est pas résolvable pour tout second membre. Si pour le second membre considéré existe une solution, elle n'est pas unique et est déterminée à la précision du terme Y^* près. Une des solutions possibles se dégage au stade de l'inversion de la matrice dégénérée $B^{(n)}$. La situation mentionnée a lieu lors de la résolution du problème de Neumann sur l'équation de Poisson dans un rectangle. Ces questions seront traitées en plus de détails au chapitre XII consacré à la résolution des équations de mailles dégénérées.

2. Problème périodique. Les problèmes vectoriels tripouctuels périodiques apparaissent avec la résolution des équations elliptiques par des méthodes des différences finies dans des systèmes des coordonnées curvilignes orthogonales: cylindrique, polaire et sphérique. Au point 3, § 1 on a donné des exemples de problèmes différentiels dont les schémas aux différences peuvent être réduits au problème suivant: rechercher la solution des équations

$$\begin{aligned} -Y_{j-1} + CY_j - Y_{j+1} = F_j, \quad & 1 \leqslant j \leqslant N-1, \\ -Y_{N-1} + CY_0 - Y_1 = F_0, \quad & j = 0, \quad Y_N = Y_0. \end{aligned} \tag{21}$$

Le problème (21) peut également être résolu à l'aide de la méthode de réduction totale. Etudions la première opération du procédé d'élimination des inconnues. Comme auparavant, des équations du système (21), pour $j = 2, 4, 6, \ldots, N-2$, éliminons à l'aide des deux équations voisines les inconnues Y_j aux numéros j impairs. Il vient

$$-Y_{j-2} + C^{(1)}Y_j - Y_{j+2} = F_j^{(1)}, \quad j = 2, 4, 6, \ldots, N-2. \tag{22}$$

Il reste à éliminer Y_1 et Y_{N-1} des équations (21) pour $j = 0$. A cette fin écrivons les trois équations suivantes du système (21):

$$\begin{aligned} -Y_0 + CY_1 - Y_2 = F_1, \quad & j = 1, \\ -Y_{N-1} + CY_0 - Y_1 = F_0, \quad & j = 0, \\ -Y_{N-2} + CY_{N-1} - Y_N = F_{N-1}, \quad & j = N-1, \end{aligned}$$

multiplions la seconde équation à gauche par C et additionnons les trois équations, compte tenu de ce que $Y_N = Y_0$. On obtient finalement

$$-Y_{N-2} + C^{(1)}Y_0 - Y_2 = F_0^{(1)}, \quad Y_N = Y_0, \tag{23}$$

où

$$F_0^{(1)} = F_1^{(0)} + C^{(0)}F_0^{(0)} + F_{N-1}^{(0)}, \quad C^{(0)} = C, \quad F_j^{(0)} \equiv F_j \; .$$

En réunissant (22) et (23), on obtient le système complet d'inconnues Y_j aux numéros j pairs, dont la structure est analogue à (21). Les inconnues Y_j aux numéros j impairs s'obtiennent à partir des équations ordinaires

$$C^{(0)}Y_j = F_j^{(0)} + Y_{j-1} + Y_{j+1}, \quad j = 1, 3, 5, \ldots, N-1.$$

Le procédé d'élimination peut être poursuivi plus loin. Après la l-ième opération d'élimination, on obtient le système d'inconnues Y_j aux numéros j multiples de 2^l:

$$\begin{aligned} -Y_{j-2^l} + C^{(l)}Y_j - Y_{j+2^l} = F_j^{(l)}, \quad & j = 2^l, 2\cdot 2^l, 3\cdot 2^l, \ldots, N-2^l, \\ -Y_{N-2^l} + C^{(l)}Y_0 - Y_{2^l} = F_0^{(l)}, \quad & j = 0, \quad Y_N = Y_0, \end{aligned}$$

et le groupe d'équations

$$C^{(k-1)}Y_j = F_j^{(k-1)} + Y_{j-2^{k-1}} + Y_{j+2^{k-1}},$$
$$j = 2^{k-1},\ 3\cdot 2^{k-1},\ 5\cdot 2^{k-1}, \ldots, N-2^{k-1}, \quad k = l,\ l-1, \ldots, 1 \tag{24}$$

permettant d'obtenir successivement les inconnues restantes. Les seconds membres $F_j^{(k)}$ s'obtiennent par récurrence pour $k = 1, 2, \ldots, n-1$:

$$F_j^{(k)} = F_{j-2^{k-1}}^{(k-1)} + C^{(k-1)}F_j^{(k-1)} + F_{j+2^{k-1}}^{(k-1)},$$
$$j = 2^k,\ 2\cdot 2^k,\ 3\cdot 2^k, \ldots, N-2^k, \tag{25}$$
$$F_0^{(k)} = F_{2^{k-1}}^{(k-1)} + C^{(k-1)}F_0^{(k-1)} + F_{N-2^{k-1}}^{(k-1)}, \quad F_j^{(0)} \equiv F_j.$$

Après la $(n-1)$-ième opération d'élimination, on obtient le système en Y_0 et $Y_{2^{n-1}}$ $(Y_N = Y_0)$:

$$\begin{aligned} C^{(n-1)}Y_0 - 2Y_{2^{n-1}} &= F_0^{(n-1)}, \\ -2Y_0 + C^{(n-1)}Y_{2^{n-1}} &= F_{2^{n-1}}^{(n-1)}. \end{aligned} \tag{26}$$

En résolvant ce système, on trouve Y_0, $Y_{2^{n-1}}$ et $Y_N = Y_0$, quant aux autres inconnues, en vertu de (24), elles seront obtenues après la résolution des équations

$$C^{(k-1)}Y_j = F_j^{(k-1)} + Y_{j-2^{k-1}} + Y_{j+2^{k-1}},$$
$$j = 2^{k-1},\ 3\cdot 2^{k-1},\ 5\cdot 2^{k-1}, \ldots, N-2^{k-1}, \quad k = n-1,\ n-2, \ldots, 1.$$

Avant de passer à la résolution de (26), cherchons les formules de récurrence pour les vecteurs $p_j^{(k)}$ et $q_j^{(k)}$ associés à $F_j^{(k)}$ par la relation suivante:

$$F_j^{(k)} = C^{(k)}p_j^{(k)} + q_j^{(k)}, \quad j = 0,\ 2^k,\ 2\cdot 2^k,\ 3\cdot 2^k, \ldots, N-2^k.$$

En utilisant les formules de récurrence (25) pour $F_j^{(k)}$, il vient

$$\begin{aligned} &C^{(k-1)}S_j^{(k-1)} = q_j^{(k-1)} + p_{j-2^{k-1}}^{(k-1)} + p_{j+2^{k-1}}^{(k-1)}, \\ &p_j^{(k)} = p_j^{(k-1)} + S_j^{(k-1)}, \\ &q_j^{(k)} = 2p_j^{(k)} + q_{j-2^{k-1}}^{(k-1)} + q_{j+2^{k-1}}^{(k-1)}, \\ &j = 2^k,\ 2\cdot 2^k,\ 3\cdot 2^k, \ldots, N-2^k, \quad k = 1, 2, \ldots, n-1, \\ &q_j^{(0)} \equiv F_j, \quad p_j^{(0)} \equiv 0, \quad j = 1, 2, \ldots, N-1, \end{aligned} \tag{27}$$

à partir desquelles on tire $p_j^{(k)}$ et $q_j^{(k)}$ pour $j \neq 0$, ainsi que les formules

$$\begin{aligned} &C^{(k-1)}S_0^{(k-1)} = q_0^{(k-1)} + p_{2^{k-1}}^{(k-1)} + p_{N-2^{k-1}}^{(k-1)}, \\ &p_0^{(k)} = p_0^{(k-1)} + S_0^{(k-1)}, \\ &q_0^{(k)} = 2p_0^{(k)} + q_{2^{k-1}}^{(k-1)} + q_{N-2^{k-1}}^{(k-1)}, \quad k = 1, 2, \ldots, n-1, \\ &q_0^{(0)} = F_0, \quad p_0^{(0)} = 0 \end{aligned} \tag{28}$$

permettant de trouver $p_0^{(k)}$ et $q_0^{(k)}$.

Abordons maintenant la résolution du système (26). A partir de (27) et (28) pour $k = n - 1$, on obtient les relations

$$q_{2^{n-1}}^{(n-1)} = 2p_{2^{n-1}}^{(n-1)} + q_{2^{n-2}}^{(n-2)} + q_{3 \cdot 2^{n-2}}^{(n-2)},$$

$$q_0^{(n-1)} = 2p_0^{(n-1)} + q_{2^{n-2}}^{(n-2)} + q_{3 \cdot 2^{n-2}}^{(n-2)},$$

à partir desquelles on tire

$$q_0^{(n-1)} - q_{2^{n-1}}^{(n-1)} = 2\left(p_0^{(n-1)} - p_{2^{n-1}}^{(n-1)}\right). \tag{29}$$

Ôtons maintenant de la première équation du système (26) la seconde équation. On obtient compte tenu de (29) et de l'égalité (12) du point 1, § 2

$$\begin{aligned} (C^{(n-1)} + 2E)\,(Y_0 - Y_{2^{n-1}}) &= \\ &= [C^{(n-2)}]^2\,(Y_0 - Y_{2^{n-1}}) = F_0^{(n-1)} - F_{2^{n-1}}^{(n-1)} = \\ &= C^{(n-1)}\left(p_0^{(n-1)} - p_{2^{n-1}}^{(n-1)}\right) + q_0^{(n-1)} - q_{2^{n-1}}^{(n-1)} = \\ &= [C^{(n-2)}]^2\left(p_0^{(n-1)} - p_{2^{n-1}}^{(n-1)}\right). \end{aligned}$$

En posant par hypothèse que $C^{(n-2)}$ est une matrice non dégénérée, on obtient de cette expression la relation

$$Y_{2^{n-1}} = Y_0 - p_0^{(n-1)} + p_{2^{n-1}}^{(n-1)}. \tag{30}$$

En portant (30) dans la première équation du système (26), il vient

$$\begin{aligned} (C^{(n-1)} - 2E)\,Y_0 &= F_0^{(n-1)} - 2\left(p_0^{(n-1)} - p_{2^{n-1}}^{(n-1)}\right) = \\ &= (C^{(n-1)} - 2E)\,p_0^{(n-1)} + q_0^{(n-1)} + 2p_{2^{n-1}}^{(n-1)}. \end{aligned}$$

Donc Y_0 peut être obtenu suivant les formules

$$\begin{aligned} &B^{(n-1)}t^{(n-1)} = q_0^{(n-1)} + 2p_{2^{n-1}}^{(n-1)}, \quad B^{(n-1)} = C^{(n-1)} - 2E, \\ &Y_0 = p_0^{(n-1)} + t^{(n-1)}, \end{aligned} \tag{31}$$

tandis que $Y_{2^{n-1}}$ pourra également être déterminé, en vertu de (30), de la relation

$$Y_{2^{n-1}} = p_{2^{n-1}}^{(n-1)} + t^{(n-1)}. \tag{32}$$

Les autres inconnues s'obtiendront successivement à l'aide des formules

$$\begin{gathered} Y_N = Y_0, \\ C^{(k-1)} t_j^{(k-1)} = q_j^{(k-1)} + Y_{j-2^{k-1}} + Y_{j+2^{k-1}}, \\ Y_j = p_j^{(k-1)} + t_j^{(k-1)}, \\ j = 2^{k-1},\ 3\cdot 2^{k-1},\ 5\cdot 2^{k-1}, \dots, N - 2^{k-1}, \\ k = n-1,\ n-2, \dots, 1. \end{gathered} \tag{33}$$

En résumé, les formules (27), (28), (31)-(33) décrivent la méthode de réduction totale appliquée à la résolution du problème périodique (21). Pour l'inversion des matrices $C^{(k-1)}$ et $B^{(n-1)}$ on utilise les factorisations (19), (20), mais dans (20) il faut substituer $n - 1$ à n.

Donnons l'estimation du nombre d'opérations arithmétiques Q que coûte la mise en œuvre de la méthode de réduction totale au cas d'un problème périodique. Désignons, comme auparavant, par $\overset{\circ}{q}$ le nombre d'opérations utilisées pour la résolution de l'équation $C_{l,k-1} V = F$, et par $\bar{q}$ celui d'opérations auxiliaires mises en œuvre à la résolution de la même équation, mais avec un autre second membre F. L'estimation s'obtient à l'aide de la formule

$$Q = \bar{q} N \log_2 N + (1{,}5\overset{\circ}{q} - 2\bar{q} + 7M) N - 2\overset{\circ}{q} + 2\bar{q} - 14M.$$

La comparaison de cette estimation à l'estimation (45) du § 2, obtenue pour le premier problème aux limites, montre que, pratiquement, le problème périodique exige le même nombre d'opérations que le premier problème aux limites.

3. Troisième problème aux limites.

3.1. *Procédé d'élimination des inconnues.* Examinons maintenant la méthode de réduction totale appliquée à la résolution du troisième problème aux limites pour équations vectorielles triponctuelles

$$\begin{aligned} (C + 2\alpha E) Y_0 - 2Y_1 &= F_0, && j = 0, \\ -Y_{j-1} + CY_j - Y_{j+1} &= F_j, && 1 \leqslant j \leqslant N-1, \\ -2Y_{N-1} + (C + 2\beta E) Y_N &= F_N, && j = N. \end{aligned} \tag{34}$$

En posant que les conditions $\alpha \geqslant 0$, $\beta \geqslant 0$, $\alpha^2 + \beta^2 \neq 0$ sont satisfaites, introduisons les notations suivantes

$$C^{(0)} = C, \quad C_1^{(0)} = C + 2\alpha E, \quad C_2^{(0)} = C + 2\beta E, \quad F_j^{(0)} \equiv F_j,$$

qui, une fois utilisées, permettent d'écrire (34) sous la forme

$$\begin{aligned} &C_1^{(0)}Y_0-2Y_1=F_0^{(0)}, && j=0,\\ &-Y_{j-1}+C^{(0)}Y_j-Y_{j+1}=F_j^{(0)}, && 1\leqslant j\leqslant N-1,\\ &-2Y_{N-1}+C_2^{(0)}Y_N=F_N^{(0)}, && j=N. \end{aligned} \tag{34'}$$

Soit $N = 2^n$. Le procédé d'élimination des inconnues pour (34') est mis en œuvre de la même façon que pour le système (1) qui correspond au cas de $C_1^{(0)} = C_2^{(0)} = C^{(0)}$ $(\alpha = \beta = 0)$.

Ecrivons le système réduit obtenu après la n-ième opération du procédé d'élimination des inconnues

$$C_1^{(n)}Y_0-2Y_N=F_0^{(n)}, \quad -2Y_0+C_2^{(n)}Y_N=F_N^{(n)}, \tag{6'}$$

ainsi que les groupes d'équations

$$\begin{gathered} C^{(k-1)}Y_j=F_j^{(k-1)}+Y_{j-2^{k-1}}+Y_{j+2^{k-1}},\\ j=2^{k-1},\ 3\cdot 2^{k-1},\ \ldots,\ N-2^{k-1},\quad k=n,\ n-1,\ \ldots,\ 1 \end{gathered} \tag{35}$$

permettant d'obtenir successivement les inconnues Y_j. Les seconds membres $F_j^{(k)}$ s'obtiennent dans ce cas à l'aide de relations de récurrence

$$\begin{gathered} F_j^{(k)}=F_{j-2^{k-1}}^{(k-1)}+C^{(k-1)}F_j^{(k-1)}+F_{j+2^{k-1}}^{(k-1)},\\ j=2^k,\ 2\cdot 2^k,\ \ldots,\ N-2^k,\quad k=1,\ 2,\ \ldots,\ n-1, \end{gathered} \tag{36}$$

$$F_0^{(k)}=C^{(k-1)}F_0^{(k-1)}+2F_{2^{k-1}}^{(k-1)}, \quad k=1,\ 2,\ \ldots,\ n, \tag{37}$$

$$F_N^{(k)}=2F_{N-2^{k-1}}^{(k-1)}+C^{(k-1)}F_N^{(k-1)}, \quad k=1,\ 2,\ \ldots,\ n, \tag{38}$$

tandis que les matrices $C_1^{(k)}$, $C_2^{(k)}$ et $C^{(k)}$ suivant les formules:

$$\begin{aligned} &C^{(k)}=[C^{(k-1)}]^2-2E, && k=1,\ 2,\ \ldots,\ n-1, && C_1^{(0)}=C,\\ &C_1^{(k)}=C^{(k-1)}C_1^{(k-1)}-2E, && k=1,\ 2,\ \ldots,\ n, && C_1^{(0)}=C+2\alpha E,\\ &C_2^{(k)}=C^{(k-1)}C_2^{(k-1)}-2E, && k=1,\ 2,\ \ldots,\ n, && C_2^{(0)}=C+2\beta E. \end{aligned} \tag{39}$$

A partir du système (6') on obtient les équations permettant de déterminer Y_0 et Y_N. De (39) on peut déduire que $C_1^{(k)}$, $C_2^{(k)}$ et $C^{(k)}$ sont des polynômes matriciels de degré 2^k relativement à la même matrice C. Ils sont donc permutables. Aussi à partir de (6') obtient-on les équations

$$\mathscr{D}^{(n+1)}Y_0=F_0^{(n+1)}, \quad C_2^{(n)}Y_N=F_N^{(n)}+2Y_0 \tag{40}$$

ainsi que des équations qui leur sont équivalentes

$$\mathscr{D}^{(n+1)}Y_N=F_N^{(n+1)}, \quad C_1^{(n)}Y_0=F_0^{(n)}+2Y_N, \tag{40'}$$

aux notations

$$F_0^{(n+1)} = C_2^{(n)} F_0^{(n)} + 2F_N^{(n)}, \tag{41}$$

$$F_N^{(n+1)} = 2F_0^{(n)} + C_1^{(n)} F_N^{(n)}, \tag{42}$$

$$\mathscr{D}^{(n+1)} = C_1^{(n)} C_2^{(n)} - 4E = C_2^{(n)} C_1^{(n)} - 4E. \tag{43}$$

Bref, pour trouver Y_0 et Y_N, on peut utiliser les équations (40) ou (40′). Utilisons (40).

Au lieu des vecteurs $F_j^{(k)}$ déterminons les vecteurs $p_j^{(k)}$ et $q_j^{(k)}$ associés à $F_j^{(k)}$ par les relations suivantes:

$$F_0^{(k)} = C_1^{(k)} p_0^{(k)} + q_0^{(k)}, \tag{44}$$

$$F_N^{(k)} = C_2^{(k)} p_N^{(k)} + q_N^{(k)}, \quad k = 0, 1, \ldots, n, \tag{45}$$

$$F_0^{(n+1)} = \mathscr{D}^{(n+1)} p_0^{(n+1)} + q_0^{(n+1)}, \tag{46}$$

$$F_j^{(k)} = C^{(k)} p_j^{(k)} + q_j^{(k)}, \tag{47}$$

$$j = 2^k,\ 2\cdot 2^k,\ \ldots,\ N - 2^k,\ k = 0,\ 1,\ 2,\ \ldots,\ n-1.$$

Cherchons les formules de récurrence pour $p_j^{(k)}$ et $q_j^{(k)}$. Si $j \neq 0$, N, on obtiendra de (36), (39) et (47), en faisant l'hypothèse, comme auparavant, sur la non-dégénérescence des matrices $C^{(k-1)}$, les formules suivantes:

$$\begin{aligned} &C^{(k-1)} S_j^{(k-1)} = q_j^{(k-1)} + p_{j-2^{k-1}}^{(k-1)} + p_{j+2^{k-1}}^{(k-1)}, \\ &p_j^{(k)} = p_j^{(k-1)} + S_j^{(k-1)}, \\ &q_j^{(k)} = 2p_j^{(k-1)} + q_{j-2^{k-1}}^{(k-1)} + q_{j+2^{k-1}}^{(k-1)}, \\ &j = 2^k,\ 2\cdot 2^k,\ \ldots,\ N - 2^k, \quad k = 1,\ 2,\ \ldots,\ n-1, \\ &q_j^{(0)} \equiv F_j, \quad p_j^{(0)} \equiv 0. \end{aligned} \tag{48}$$

Cherchons les formules pour $p_0^{(k)}$ et $q_0^{(k)}$ avec $k = 0, 1, \ldots, n+1$. En portant (44) et (47) dans (37) et (44)-(46) dans (41), on obtient pour $k = 1, 2, \ldots, n$

$$C_1^{(k)} p_0^{(k)} + q_0^{(k)} = C^{(k-1)} \left(C_1^{(k-1)} p_0^{(k-1)} + q_0^{(k-1)} + 2p_{2^{k-1}}^{(k-1)}\right) + 2q_{2^{k-1}}^{(k-1)} \tag{49}$$

et pour $k = n+1$

$$\mathscr{D}^{(n+1)} p_0^{(n+1)} + q_0^{(n+1)} = C_2^{(n)} \left(C_1^{(n)} p_0^{(n)} + q_0^{(n)} + 2p_N^{(n)}\right) + 2q_N^{(n)}. \tag{50}$$

Choisissons $q_0^{(k)}$ et $q_0^{(n+1)}$ suivant les formules

$$\begin{aligned} &q_0^{(k)} = 2p_0^{(k)} + 2q_{2^{k-1}}^{(k-1)}, \quad k = 1,\ 2,\ \ldots,\ n, \\ &q_0^{(n+1)} = 4p_0^{(n+1)} + 2q_N^{(n)} \end{aligned} \tag{51}$$

et utilisons les égaités découlant de (39) et (43)

$$C_1^{(k)}+2E=C^{(k-1)}C_1^{(k-1)}, \quad \mathcal{D}^{(n+1)}+4E=C_2^{(n)}C_1^{(n)}.$$

Dans ce cas on peut écrire (49) et (50), sous l'hypothèse de la non-dégénérescence de $C^{(k-1)}$ et $C_2^{(n)}$, en forme d'une équation unique

$$C_1^{(k-1)}p_0^{(k)}=C_1^{(k-1)}p_0^{(k-1)}+q_0^{(k-1)}+2p_{2^{k-1}}^{(k-1)},$$
$$k=1,\ 2,\ \ldots,\ n+1.$$

En joignant ces équations à (51), on obtient finalement les formules pour le calcul de $p_0^{(k)}$ et $q_0^{(k)}$:

$$\begin{aligned}
&C_1^{(k-1)}S_0^{(k-1)}=q_0^{(k-1)}+2p_{2^{k-1}}^{(k-1)},\\
&p_0^{(k)}=p_0^{(k-1)}+S_0^{(k-1)}, \quad k=1,\ 2,\ \ldots,\ n+1,\\
&q_0^{(k)}=2p_0^{(k)}+2q_{2^{k-1}}^{(k-1)}, \quad k=1,\ 2,\ \ldots,\ n,\\
&q_0^{(n+1)}=4p_0^{(n+1)}+2q_N^{(n)},\\
&q_0^{(0)}=F_0, \quad p_0^{(0)}=0.
\end{aligned} \tag{52}$$

De façon analogue, en utilisant (45), (47), les relations de récurrence (38) et (39), on obtient les formules pour le calcul de $p_N^{(k)}$ et $q_N^{(k)}$:

$$\begin{aligned}
&C_2^{(k-1)}S_N^{(k-1)}=q_N^{(k-1)}+2p_{N-2^{k-1}}^{(k-1)},\\
&p_N^{(k)}=p_N^{(k-1)}+S_N^{(k-1)},\\
&q_N^{(k)}=2p_N^{(k)}+2q_{N-2^{k-1}}^{(k-1)}, \quad k=1,\ 2,\ \ldots,\ n,\\
&q_N^{(0)}=F_N, \quad p_N^{(0)}=0.
\end{aligned} \tag{53}$$

Il reste à éliminer $F_j^{(k)}$ de (35) et (40). En portant (47) dans (35) et (45) et (46) dans (40), on obtient les formules suivantes permettant de trouver Y_j:

$$\mathcal{D}^{(n+1)}S_0^{(n+1)}=q_0^{(n+1)}, \quad Y_0=p_0^{(n+1)}+S_0^{(n+1)}, \tag{54}$$

$$C_2^{(n)}S_N^{(n)}=q_N^{(n)}+2Y_0, \quad Y_N=p_N^{(n)}+S_N^{(n)}, \tag{55}$$

$$\begin{aligned}
&C^{(k-1)}S_j^{(k-1)}=q_j^{(k-1)}+Y_{j-2^{k-1}}+Y_{j+2^{k-1}},\\
&Y_j=p_j^{(k-1)}+S_j^{(k-1)},
\end{aligned} \tag{56}$$

$$j=2^{k-1},\ 3\cdot 2^{k-1},\ \ldots,\ N-2^{k-1}, \quad k=n,\ n-1,\ \ldots,\ 1.$$

En résumé, les formules (48), (52)-(56) décrivent la méthode de réduction totale appliquée à la résolution du troisième problème aux limites (34).

R e m a r q u e 1. Si pour trouver Y_0 et Y_N on a utilisé les équations (40′), alors, en introduisant au lieu de $p_0^{(n+1)}$ et $q_0^{(n+1)}$ les vecteurs $p_N^{(n+1)}$ et $q_N^{(n+1)}$ associés à $F_N^{(n+1)}$ par la relation

$$F_N^{(n+1)} = \mathcal{Q}^{(n+1)} p_N^{(n+1)} + q_N^{(n+1)},$$

on obtient à partir de (38), (42), (44) et (47) les formules suivantes permettant de trouver $p_N^{(k)}$ et $q_N^{(k)}$:

$$\begin{aligned} &C_2^{(k-1)} S_N^{(k-1)} = q_N^{(k-1)} + 2p_{N-2^{k-1}}^{(k-1)}, \\ &p_N^{(k)} = p_N^{(k-1)} + S_N^{(k-1)}, \qquad k = 1,\ 2,\ \ldots,\ n+1, \\ &q_N^{(k)} = 2p_N^{(k)} + 2q_{N-2^{k-1}}^{(k-1)}, \quad k = 1,\ 2,\ \ldots,\ n, \\ &q_N^{(n+1)} = 4p_N^{(n+1)} + 2q_0^{(n)}, \\ &q_N^{(0)} = F_N, \quad p_N^{(0)} = 0. \end{aligned} \tag{53′}$$

Les formules (53′) remplacent les formules (53). Vu que dans ce cas on n'est pas obligé de calculer le vecteur $F_0^{(n+1)}$ et, par suite, les vecteurs $p_0^{(n+1)}$ et $q_0^{(n+1)}$, les formules (52) sont remplacées par

$$\begin{aligned} &C_1^{(k-1)} S_0^{(k-1)} = q_0^{(k-1)} + 2p_{2^{k-1}}^{(k-1)}, \quad p_0^{(k)} = p_0^{(k-1)} + S_0^{(k-1)}, \\ &q_0^{(k)} = 2p_0^{(k)} + 2q_{2^{k-1}}^{(k-1)}, \quad k = 1,\ 2,\ \ldots,\ n, \\ &q_0^{(0)} = F_0, \quad p_0^{(0)} = 0. \end{aligned} \tag{52′}$$

A partir de (35) et (40′) on obtient les formules pour l'obtention de Y_0 et Y_N:

$$\mathcal{Q}^{(n+1)} S_N^{(n+1)} = q_N^{(n+1)}, \quad Y_N = p_N^{(n+1)} + S_N^{(n+1)}, \tag{55′}$$

$$C_1^{(n)} S_0^{(n)} = q_0^{(n)} + 2Y_N, \quad Y_0 = p_0^{(n)} + S_0^{(n)}. \tag{54′}$$

Les inconnues restantes s'obtiennent à l'aide de (56). Les formules (48), (52′)-(55′) et (56) peuvent également être utilisées à la résolution du problème (34).

R e m a r q u e 2. Si Y_N est donné, c'est-à-dire si au lieu de (34) il faut résoudre le problème aux limites

$$\begin{aligned} (C + 2\alpha E)\, Y_0 - 2Y_1 &= F_0, \qquad j = 0, \\ -Y_{j-1} + CY_j - Y_{j+1} &= F_j, \quad 1 \leqslant j \leqslant N-1, \\ Y_N &= F_N, \qquad j = N, \end{aligned}$$

alors, dans ce cas, la méthode de réduction totale se décrit par les formules (48), (52′), (54′) et (56). Si, par contre, est donné Y_0, c'est-

à-dire s'il s'agit de résoudre le problème

$$-Y_{j-1} + CY_j - Y_{j+1} = F_j, \quad 1 \leqslant j \leqslant N-1,$$
$$-2Y_{N-1} + (C + 2\beta E)\, Y_N = F_N, \quad j = N,\ Y_0 = F_0,$$

la méthode est alors décrite par les formules (48), (53), (55) et (56).

3.2. *Factorisation des matrices.* Il s'ensuit de (39) et (43) que $C_1^{(k)}$, $C_2^{(k)}$ et $C^{(k)}$ sont des polynômes matriciels de degré 2^k, tandis que $\mathscr{D}^{(n+1)}$ l'est de degré 2^{n+1} relativement à la matrice C avec coefficient 1 près du degré majeur. Vu la nécessité d'invertir ces matrices, procédons à leur factorisation. A cette fin cherchons la représentation explicite de ces polynômes au moyen des polynômes connus et étudions le problème de l'obtention des racines des polynômes considérés.

On a montré au point 2 du § 2 que $C^{(k)}$ s'expriment au moyen du polynôme de Tchébychev de première espèce de la façon suivante:

$$C^{(k)} = 2T_{2^k}\left(\frac{1}{2}C\right), \quad k = 0, 1, \ldots \tag{57}$$

Ensuite, de la relation (39) on tire:

$$C_1^{(k)} - C^{(k)} = C^{(k-1)}[C_1^{(k-1)} - C^{(k-1)}] = \ldots$$
$$\ldots = \prod_{l=0}^{k-1} C^{(l)}[C_1^{(0)} - C^{(0)}] = 2\alpha \prod_{l=0}^{k-1} C^{(l)}. \tag{58}$$

Vu qu'il y a lieu l'égalité

$$\prod_{l=0}^{k-1} C^{(l)} = \prod_{l=0}^{k-1} 2T_{2^l}\left(\frac{1}{2}C\right) = U_{2^k-1}\left(\frac{1}{2}C\right),$$

où $U_n(x)$ est le polynôme de Tchébychev de seconde espèce, on obtient de (58) les représentations suivantes pour $C_1^{(k)}$:

$$C_1^{(k)} = 2T_{2^k}\left(\frac{1}{2}C\right) + 2\alpha U_{2^k-1}\left(\frac{1}{2}C\right), \quad k = 0, 1, \ldots \tag{59}$$

De façon analogue on obtient la représentation de $C_2^{(k)}$:

$$C_2^{(k)} = 2T_{2^k}\left(\frac{1}{2}C\right) + 2\beta U_{2^k-1}\left(\frac{1}{2}C\right), \quad k = 0, 1, \ldots \tag{60}$$

Ensuite, en portant (59) et (60) dans (43), il vient

$$\mathscr{D}^{(n+1)} = 4\left[T_{2^k}\left(\frac{1}{2}C\right)\right]^2 - 4E +$$
$$+ 4(\alpha+\beta)\, T_{2^k}\left(\frac{1}{2}C\right) U_{2^k-1}\left(\frac{1}{2}C\right) + 4\alpha\beta\left[U_{2^k-1}\left(\frac{1}{2}C\right)\right]^2. \tag{61}$$

Vu qu'il y a lieu l'égalité

$$1-T_n(x)=U_{n-1}(x)(1-x^2), \qquad (62)$$

il s'ensuit de (61)

$$\mathscr{D}^{(n+1)}=U_{2^n-1}\left(\frac{1}{2}C\right)\left[(C^2+4\alpha\beta E-4E)U_{2^n-1}\left(\frac{1}{2}C\right)+\right.$$
$$\left.+4(\alpha+\beta)T_{2^n}\left(\frac{1}{2}C\right)\right].$$

Bref, on a obtenu la représentation de $C^{(k)}$, $C_1^{(k)}$, $C_2^{(k)}$ et $\mathscr{D}^{(n+1)}$ au moyen des polynômes connus. Vu que les racines des polynômes de Tchébychev de première et de seconde espèces sont connues, on tire de (57) et (62)

$$C^{(k)}=\sum_{l=1}^{2^k}\left(C-2\cos\frac{(2l-1)\pi}{2^{k+1}}E\right),$$

$$\mathscr{D}^{(n+1)}=\sum_{l=1}^{2^n-1}\left(C-2\cos\frac{l\pi}{2^n}E\right)\left[(C^2+4\alpha\beta E-4E)U_{2^n-1}\left(\frac{1}{2}C\right)+\right.$$
$$\left.+4(\alpha+\beta)T_{2^n}\left(\frac{1}{2}C\right)\right].$$

Aussi s'ensuit-il de (59), (60) qu'il ne reste qu'à trouver les racines des polynômes

$$\begin{gathered}P_m(t)=2T_m\left(\frac{t}{2}\right)+2\alpha U_{m-1}\left(\frac{t}{2}\right),\\ Q_m(t)=2T_m\left(\frac{t}{2}\right)+2\beta U_{m-1}\left(\frac{t}{2}\right),\\ m=2^k,\quad k=0,\ 1,\ \ldots,\ n-1,\end{gathered} \qquad (63)$$

qui correspondent aux polynômes matriciels $C_1^{(k)}$ et $C_2^{(k)}$ et les racines du polynôme

$$R_{2^n+1}(t)=(t^2+4\alpha\beta-4)U_{2^n-1}\left(\frac{t}{2}\right)+4(\alpha+\beta)T_{2^n}\left(\frac{t}{2}\right), \qquad (64)$$

qui engendre le polynôme $\mathscr{D}^{(n+1)}$.

Ce problème peut être résolu de deux façons. La première consiste à utiliser l'une des méthodes d'obtention approchée des racines du polynôme, la seconde dans la réduction de ce problème à celui de l'obtention de toutes les valeurs propres de certaines matrices tridiagonales. Arrêtons-nous en plus de détails sur le second procédé.

Désignons par $S_k(\lambda)$ le déterminant de k-ième ordre suivant:

$$S_k(\lambda)=\begin{vmatrix}\lambda+2\alpha & 2 & 0 & 0 & \ldots & 0 & 0 & 0 & 0\\ 1 & \lambda & 1 & 0 & \ldots & 0 & 0 & 0 & 0\\ 0 & 1 & \lambda & 1 & \ldots & 0 & 0 & 0 & 0\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & 0 & 0 & \ldots & 1 & \lambda & 1 & 0\\ 0 & 0 & 0 & 0 & \ldots & 0 & 1 & \lambda & 1\\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 1 & \lambda\end{vmatrix}, \quad k\geqslant 2$$

et posons $S_1(\lambda) = \lambda + 2\alpha$. A partir de la structure et de la définition de la matrice correspondant à $S_k(\lambda)$, on déduit les relations de récurrence pour $S_k(\lambda)$:

$$\begin{aligned} &S_{k+1}(\lambda) = \lambda S_k(\lambda) - S_{k-1}(\lambda), \quad k \geqslant 2, \\ &S_2(\lambda) = \lambda S_1(\lambda) - 2, \quad S_1(\lambda) = \lambda + 2\alpha. \end{aligned} \tag{65}$$

En utilisant les relations de récurrence pour le polynôme de Tchébychev (voir point 2, § 4, ch. I)

$$\begin{aligned} &T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad T_1(x) = x, \quad T_0(x) = 1, \\ &U_{n+1}(x) = 2xU_n(x) - U_{n-1}(x), \quad U_1(x) = 2x, \quad U_0(x) = 1 \end{aligned}$$

et les relations (65), on obtient la représentation de $S_m(\lambda)$ au moyen des polynômes de Tchébychev: $S_m(\lambda) = 2T_m\left(\frac{\lambda}{2}\right) + 2\alpha U_{m-1}\left(\frac{\lambda}{2}\right)$, $m \geqslant 1$. En comparant cette expression à (63), on trouve que les racines du polynôme $P_m(t)$ coïncident avec celles du déterminant $S_m(\lambda)$ dépendant de λ à la façon d'un paramètre.

Le problème de la recherche des racines de $S_m(\lambda)$ est équivalent à celui de la recherche de telles valeurs du paramètre λ pour lesquelles le système d'équations algébriques

$$\begin{aligned} &y_{i-1} + \lambda y_i + y_{i+1} = 0, \quad 1 \leqslant i \leqslant m-1, \\ &(\lambda + 2\alpha)\, y_0 + 2y_1 = 0, \qquad i = 0, \\ &y_m = 0 \end{aligned} \tag{66}$$

possède une solution non nulle. Donnons une autre transcription de (66). Utilisons la notation de la différence divisée seconde

$$y_{\bar{x}x,\, i} = \frac{1}{h}\left(y_{x,\, i} - y_{\bar{x},\, i}\right) = \frac{1}{h^2}\left(y_{i+1} - 2y_i + y_{i-1}\right)$$

et récrivons (66) sous la forme suivante:

$$\begin{aligned} &y_{\bar{x}x} + \mu y = 0, \quad 1 \leqslant i \leqslant m-1, \\ &\frac{2}{h} y_x + \frac{2\alpha}{h^2} y + \mu y = 0, \qquad i = 0, \quad y_m = 0, \end{aligned} \tag{66'}$$

où λ et μ sont associés par la relation $\lambda = \mu h^2 - 2$. En résumé, pour obtenir les racines du polynôme $C_1^{(k)}$ il suffit de résoudre le problème (66') pour $m = 2^k$, $k = 0, 1, \ldots$

Par analogie avec ce qui vient d'être exposé, on peut montrer que les racines du polynôme $Q_m(t)$ s'obtiennent en résolvant le problème

$$\begin{aligned} &y_{\bar{x}x} + \mu y = 0, \quad 1 \leqslant i \leqslant m-1, \\ &-\frac{2}{h} y_{\bar{x}} + \frac{2\beta}{h^2} y + \mu y = 0, \qquad i = m, \quad y_0 = 0, \end{aligned} \tag{67}$$

la relation $\lambda = \mu h^2 - 2$ déterminant ces racines.

Pour trouver les racines du polynôme $R_{2^n+1}(t)$ défini dans (64), il faut résoudre le problème de valeurs propres suivant:

$$\begin{aligned} &y_{\bar{x}x} + \mu y = 0, \qquad 1 \leqslant i \leqslant 2^n - 1, \\ &\frac{2}{h} y_x + \frac{2\alpha}{h^2} y + \mu y = 0, \quad i = 0, \\ &-\frac{2}{h} y_{\bar{x}} + \frac{2\beta}{h^2} y + \mu y = 0, \quad i = 2^n, \end{aligned} \tag{68}$$

et trouver les racines à partir de l'égalité $\lambda = \mu h^2 - 2$.

Notons que pour résoudre les problèmes (66)-(68) on peut utiliser l'algorithme connu QR de la résolution du problème complet des valeurs propres.

CHAPITRE IV

MÉTHODE DE SÉPARATION DES VARIABLES

On étudie dans ce chapitre les différentes variantes de la méthode de séparation des variables utilisée pour résoudre les équations de mailles elliptiques les plus simples dans un rectangle. Dans le premier paragraphe on décrit l'algorithme de la transformation discrète rapide de Fourier des fonctions réelles et complexes. Le § 2 est consacré à l'étude de la variante classique de la méthode de séparation des variables qui utilise l'algorithme de la transformation de Fourier. Dans le § 3 on construit la méthode combinée comprenant la réduction incomplète et la séparation des variables. On passe en revue les applications de cette méthode à la résolution des problèmes aux limites discrets relativement à l'équation de Poisson de second et quatrième ordre de précision.

§ 1. Algorithme de la transformation discrète de Fourier

1. Position du problème. Un des procédés de recherche des solutions aux problèmes de mailles multidimensionnels, admettant la séparation des variables, est le développement de la solution cherchée en une somme finie de Fourier suivant les fonctions propres correspondant aux opérateurs de mailles. L'efficience de la méthode est essentiellement fonction de la rapidité de calcul des coefficients de Fourier de la fonction de maille donnée et du rétablissement de la fonction cherchée à l'aide des coefficients de Fourier donnés.

Si, par exemple, sur le maillage $\bar{\omega} = \{x_i = ih,\ 0 \leqslant i \leqslant N,\ hN = l\}$ comprenant $N + 1$ nœuds sont donnés la fonction $f(i)$ et le système de fonctions orthonormées $\mu_k(i)$, $k = 0, 1, \ldots, N$, tandis que les coefficients de Fourier de la fonction $f(i)$ se calculent suivant les formules

$$\varphi_k = \sum_{i=0}^{N} f(i)\, \mu_k(i)\, h, \quad k = 0,\ 1,\ \ldots,\ N, \tag{1}$$

il suffit alors pour le calcul de tous les coefficients φ_k $(N + 1) \times (N + 2)$ opérations de multiplication et $N(N + 1)$ opérations d'addition.

Dans le cas général d'un système arbitraire de fonctions $\{\mu_k(i)\}$ c'est la quantité minimale d'opérations arithmétiques nécessaires.

Dans une série de cas particuliers, quand le système orthonormé de fonctions est d'un aspect spécial, le nombre total d'opérations arithmétiques nécessaire au calcul de la somme de la forme (1) peut être abaissé d'une façon sensible. On examinera ces cas en donnant les algorithmes permettant de calculer tous les coefficients de Fourier et de rétablir la fonction à l'aide des coefficients de Fourier donnés en $O(N \ln N)$ opérations arithmétiques.

Passons à la description des cas mentionnés.

P r o b l è m e 1. *Développement en sinus.* Soit sur le segment $0 \leqslant x \leqslant l$ le maillage régulier $\bar{\omega} = \{x_j = jh,\ 0 \leqslant j \leqslant N,\ hN = l\}$ au pas h. Désignons par $\omega = \{x_j = jh,\ 1 \leqslant j \leqslant N-1\}$ l'ensemble des nœuds internes du maillage $\bar{\omega}$.

Soit donnée sur ω la fonction de maille réelle $f(j)$ (ou $f(j)$ est donnée sur $\bar{\omega}$, avec $f(0) = f(N) = 0$).

Au § 5, ch. I on a montré que la fonction $f(j)$ peut être représentée sous forme d'un développement

$$f(j) = \frac{2}{N} \sum_{k=1}^{N-1} \varphi_k \sin \frac{k\pi j}{N}, \quad j = 1, 2, \ldots, N-1, \tag{2}$$

où les coefficients φ_k se déterminent à l'aide de la formule

$$\varphi_k = \sum_{j=1}^{N-1} f(j) \sin \frac{k\pi j}{N}, \quad k = 1, 2, \ldots, N-1. \tag{3}$$

En comparant (2) et (3), on aboutit à ce que les problèmes de calcul des coefficients φ_k de la fonction donnée $f(j)$ et de rétablissement de cette fonction au moyen des fonctions données $\{\varphi_k\}$ se ramènent au calcul de la somme $N-1$ de l'aspect

$$y_k = \sum_{j=1}^{N-1} a_j \sin \frac{k\pi j}{N}, \quad k = 1, 2, \ldots, N-1. \tag{4}$$

La formule (4) décrit la règle de transformation de la fonction de maille a_j, $1 \leqslant j \leqslant N-1$, associée au maillage ω en la fonction de maille y_j, $1 \leqslant j \leqslant N-1$. L'interprétation algébrique de (4) se représente ainsi: si l'on désigne par $a = (a_1, a_2, \ldots, a_{N-1})$ le vecteur de dimension $N-1$, (4) décrit alors la transformation du vecteur a avec le passage de la base naturelle à la base définie par le système de vecteurs orthogonaux

$$z_k = (z_k(1),\ z_k(2),\ \ldots,\ z_k(N-1)),\quad z_k(j) = \sin \frac{k\pi j}{N}.$$

P r o b l è m e 2. *Développement en sinus rapprochés.* Soit une fonction de maille $f(j)$, admettant des valeurs propres, donnée sur un ensemble $\omega^+ = \{x_j = jh,\ 1 \leqslant j \leqslant N\}$ (ou sur $\bar{\omega}$, avec $f(0) =$

$= 0$). Au ch. I, § 5 on a montré que la fonction $f(j)$ peut être représentée sous forme

$$f(j) = \frac{2}{N} \sum_{k=1}^{N} \varphi_k \sin \frac{(2k-1)\pi j}{2N}, \quad j = 1, 2, \ldots, N, \tag{5}$$

où les coefficients φ_k se déterminent suivant la formule

$$\varphi_k = \sum_{j=1}^{N} \rho_j f(j) \sin \frac{(2k-1)\pi j}{2N}, \quad k = 1, 2, \ldots, N, \tag{6}$$

tandis que

$$\rho_j = \begin{cases} 1, & j \neq 0, N; \\ 0{,}5, & j = 0, N. \end{cases} \tag{7}$$

Si la fonction $f(j)$ est donnée sur l'ensemble $\omega^- = \{x_j = jh, 0 \leqslant j \leqslant N-1\}$ (ou sur $\bar{\omega}$ avec $f(N) = 0$), alors le développement analogue à (5) et (6) prend la forme

$$f(N-j) = \frac{2}{N} \sum_{k=1}^{N} \varphi_k \sin \frac{(2k-1)\pi j}{2N}, \quad j = 1, 2, \ldots, N, \tag{8}$$

$$\varphi_k = \sum_{j=1}^{N} \rho_{N-j} f(N-j) \sin \frac{(2k-1)\pi j}{2N}, \quad k = 1, 2, \ldots, N, \tag{9}$$

où la fonction ρ_j est définie dans (7).

Il s'ensuit de (5), (6), (8) et (9) qu'on se trouve devant des problèmes de calcul des sommes de la forme

$$y_k = \sum_{j=1}^{N} a_j \sin \frac{(2k-1)\pi j}{2N}, \quad k = 1, 2, \ldots, N, \tag{10}$$

$$y_j = \sum_{k=1}^{N} a_k \sin \frac{(2k-1)\pi j}{2N}, \quad j = 1, 2, \ldots, N. \tag{10'}$$

P r o b l è m e 3. *Développement en cosinus.* Soit donnée sur le maillage $\bar{\omega}$ la fonction de maille réelle $f(j)$. On a alors pour la fonction $f(j)$ le développement

$$f(j) = \frac{2}{N} \sum_{k=0}^{N} \rho_k \varphi_k \cos \frac{k\pi j}{N}, \quad j = 0, 1, \ldots, N, \tag{11}$$

où

$$\varphi_k = \sum_{j=0}^{N} \rho_j f(j) \cos \frac{k\pi j}{N}, \quad k = 0, 1, \ldots, N, \tag{12}$$

tandis que ρ_j se détermine à partir de (7). Il s'ensuit des formules (11) et (12) le problème de calcul des sommes de la forme

$$y_k=\sum_{j=0}^{N} a_j \cos\frac{k\pi j}{N}, \quad k=0,\ 1,\ \ldots,\ N. \tag{13}$$

P r o b l è m e 4. *Transformation d'une fonction de maille périodique réelle.* Soit sur l'axe $-\infty < x < \infty$ le maillage régulier $\Omega = \{x_j = jh,\ j = 0, \pm 1, \pm 2, \ldots, Nh = l\}$ de pas h. Sur ce maillage Ω est donnée la fonction de maille périodique de période N

$$f(j) = f(j+N), \quad j = 0, \pm 1, \ldots,$$

prenant des valeurs réelles. On a montré au § 5, ch. I que la fonction $f(j)$ pour $0 \leqslant j \leqslant N-1$ peut être représentée sous forme (pour N pair)

$$f(j)=\frac{2}{N}\left[\sum_{k=0}^{N/2} \rho_k\varphi_k \cos\frac{2k\pi j}{N}+\right.$$
$$\left.+\sum_{k=1}^{N/2-1} \bar{\varphi}_k \sin\frac{2k\pi j}{N}\right], \quad j=0,\ 1,\ \ldots,\ N-1, \tag{14}$$

où les coefficients φ_k et $\bar{\varphi}_k$ se déterminent suivant les formules

$$\varphi_k=\sum_{j=0}^{N-1} f(j) \cos\frac{2k\pi j}{N}, \quad k=0,\ 1,\ \ldots,\ \frac{N}{2}, \tag{15}$$

$$\bar{\varphi}_k=\sum_{j=1}^{N-1} f(j) \sin\frac{2k\pi j}{N}, \quad k=1,\ 2,\ \ldots,\ \frac{N}{2}-1, \tag{16}$$

tandis que la fonction ρ_k vaut

$$\rho_k=\begin{cases} 1, & j\neq 0,\ N/2, \\ 0{,}5, & j=0,\ N/2. \end{cases}$$

Les formules (14)-(16) nous conduisent au problème de calcul des sommes de trois formes

$$y_k=\sum_{j=0}^{N/2} a_j \cos\frac{2k\pi j}{N}+\sum_{j=1}^{N/2-1} \bar{a}_j \sin\frac{2k\pi j}{N}, \quad k=0,\ 1,\ \ldots,\ N-1, \tag{17}$$

$$\left.\begin{aligned} y_k&=\sum_{j=0}^{N-1} a_j \cos\frac{2k\pi j}{N}, \quad k=0,\ 1,\ \ldots,\ N/2, \\ \bar{y}_k&=\sum_{j=1}^{N-1} a_j \sin\frac{2k\pi j}{N}, \quad k=1,\ 2,\ \ldots,\ N/2-1, \end{aligned}\right\} \tag{18}$$

de plus, dans les sommes (18) les coefficients a_j sont les mêmes.

Problème 5. *Transformation d'une fonction de maille périodique complexe.* Soit une fonction de maille périodique $f(j)$ de période N donnée sur le maillage Ω et qui prend maintenant des valeurs complexes. La fonction $f(j)$ pour $0 \leqslant j \leqslant N-1$ peut alors se représenter sous forme

$$f(j) = \frac{1}{N} \sum_{k=0}^{N-1} \varphi_k e^{\frac{2k\pi j}{N} i}, \quad j = 0, 1, \ldots, N-1, \quad i = \sqrt{-1}, \tag{19}$$

où les coefficients complexes φ_k se déterminent suivant la formule

$$\varphi_k = \sum_{j=0}^{N-1} f(j) e^{\frac{-2k\pi j}{N} i}, \quad k = 0, 1, \ldots, N-1. \tag{20}$$

Notons que $\varphi_0 = \varphi_N$ et, de plus,

$$\varphi_{N-k} = \sum_{j=0}^{N-1} f(j) e^{\frac{2k\pi j}{N} i}, \quad k = 0, 1, \ldots, N-1.$$

Aussi le calcul des coefficients φ_k et le rétablissement de la fonction $f(j)$ se réduisent-ils au calcul des sommes de la forme

$$y_k = \sum_{j=0}^{N-1} a_j e^{\frac{2k\pi j}{N}}, \quad k = 0, 1, \ldots, N-1 \tag{21}$$

aux a_j complexes.

Bref, il nous faut construire les algorithmes pour le calcul des sommes de la forme (4), (10), (13), (17), (18) et (21) exigeant moins que $O(N^2)$ opérations arithmétiques. La construction de l'algorithme est la plus simple quand N est une puissance de 2: $N = 2^n$; on se limitera à ce dernier cas.

2. Développement en sinus et en sinus rapprochés. Voyons en détail l'algorithme de calcul des sommes (4) en posant que $N = 2^n$. Dans ce cas (4) prend la forme

$$y_k = \sum_{j=1}^{2^n-1} a_j^{(0)} \sin \frac{k\pi j}{2^n}, \quad k = 1, 2, \ldots, 2^n - 1, \tag{22}$$

où est introduite la notation $a_j^{(0)} = a_j$.

L'idée de la méthode réside dans ce que dans la somme (22) les termes au même facteur sont groupés avant d'effectuer la multiplication. D'abord, avec la mise en œuvre de l'algorithme, on rassemble les termes de la somme (22) possédant des indices j et $2^n - j$ pour $j = 1, 2, \ldots, 2^{n-1} - 1$ en utilisant notamment l'égalité

$$\sin \frac{k\pi(2^n - j)}{2^n} = (-1)^{k-1} \sin \frac{k\pi j}{2^n}. \tag{23}$$

Ecrivons pour cela (22) sous forme de trois termes

$$y_k = \sum_{j=1}^{2^{n-1}-1} a_j^{(0)} \sin \frac{k\pi j}{2^n} + \sum_{j=2^{n-1}+1}^{2^n-1} a_j^{(0)} \sin \frac{k\pi j}{2^n} + a_{2^{n-1}}^{(0)} \sin \frac{k\pi}{2}$$

et réalisons la substitution $j' = 2^n - j$ dans la seconde somme. Compte tenu de (23), il vient

$$y_k = \sum_{j=1}^{2^{n-1}-1} [a_j^{(0)} + (-1)^{k-1} a_{2^n-j}^{(0)}] \sin \frac{k\pi j}{2^n} + a_{2^{n-1}}^{(0)} \sin \frac{k\pi}{2}. \quad (24)$$

En posant

$$a_j^{(1)} = a_j^{(0)} - a_{2^n-j}^{(0)},$$

$$a_{2^n-j}^{(1)} = a_j^{(0)} + a_{2^n-j}^{(0)}, \quad j = 1, 2, \ldots, 2^{n-1} - 1,$$

$$a_{2^{n-1}}^{(1)} = a_{2^{n-1}}^{(0)},$$

on tire de (24) que

$$y_{2k-1} = \sum_{j=1}^{2^{n-1}} a_{2^n-j}^{(1)} \sin \frac{(2k-1)\pi j}{2^n}, \quad k = 1, 2, \ldots, 2^{n-1}, \quad (25)$$

$$y_{2k} = \sum_{j=1}^{2^{n-1}-1} a_j^{(1)} \sin \frac{k\pi j}{2^{n-1}}, \quad k = 1, 2, \ldots, 2^{n-1} - 1. \quad (26)$$

Bref, après le premier pas on a deux sommes de la forme (25) et (26) dont chacune comprend environ deux fois moins de termes que la somme initiale (22). En outre, les sommes de la forme (26) et la somme initiale présentent une structure analogue. Aussi peut-on appliquer à (26) le procédé de groupement décrit plus haut.

Dans le second pas, en divisant, comme plus haut, la somme (26) en trois termes compte tenu de l'égalité (23), où à n on substitue $n-1$, on rassemble les termes de la somme (26) aux indices j et $2^{n-1} - j$ pour $j = 1, 2, \ldots, 2^{n-2} - 1$. Après ce second pas on obtient au lieu de (26)

$$y_{2(2k-1)} = \sum_{j=1}^{2^{n-2}} a_{2^{n-1}-j}^{(2)} \sin \frac{(2k-1)\pi j}{2^{n-1}}, \quad k = 1, 2, \ldots, 2^{n-2}, \quad (27)$$

$$y_{2^2k} = \sum_{j=1}^{2^{n-2}-1} a_j^{(2)} \sin \frac{k\pi j}{2^{n-2}}, \quad k = 1, 2, \ldots, 2^{n-2} - 1, \quad (28)$$

où

$$a_j^{(2)} = a_j^{(1)} - a_{2^{n-1}-j}^{(1)},$$
$$a_{2^{n-1}-j}^{(2)} = a_j^{(1)} + a_{2^{n-1}-j}^{(1)}, \quad j = 1, 2, \ldots, 2^{n-2}-1,$$
$$a_{2^{n-2}}^{(2)} = a_{2^{n-2}}^{(1)}.$$

Donc le problème initial (22) est équivalent au calcul des sommes (25), (27), (28). La formule (28) permet de calculer y_k pour des k multiples de 4, la formule (27) pour des k multiples de 2, mais non multiples de 4, et la formule (25) pour le calcul de y_k à k impair.

En continuant le procédé de transformation des sommes obtenues, on obtient finalement le résultat du p-ième pas

$$y_{2^{s-1}(2k-1)} = \sum_{j=1}^{2^{n-s}} a_{2^{n-s+1}-j}^{(s)} \sin \frac{(2k-1)\pi j}{2^{n-s+1}},$$
$$k = 1, 2, \ldots, 2^{n-s}, \quad s = 1, 2, \ldots, p, \tag{29}$$
$$y_{2^p k} = \sum_{j=1}^{2^{n-p}-1} a_j^{(p)} \sin \frac{k\pi j}{2^{n-p}}, \quad k = 1, 2, \ldots, 2^{n-p}-1,$$

où $p = 1, 2, \ldots, n-1$, quant aux coefficients $a_j^{(p)}$, on les détermine par récurrence

$$a_j^{(p)} = a_j^{(p-1)} - a_{2^{n-p+1}-j}^{(p-1)},$$
$$a_{2^{n-p+1}-j}^{(p)} = a_j^{(p-1)} + a_{2^{n-p+1}-j}^{(p-1)}, \quad j = 1, 2, \ldots, 2^{n-p}-1, \tag{30}$$
$$a_{2^{n-p}}^{(p)} = a_{2^{n-p}}^{(p-1)}, \quad p = 1, 2, \ldots, n-1.$$

En posant dans (29) $p = n-1$, il vient

$$y_{2^{n-1}} = \sum_{j=1}^{1} a_j^{(n-1)} \sin \frac{\pi j}{2} = a_1^{(n-1)},$$
$$y_{2^{s-1}(2k-1)} = \sum_{j=1}^{2^{n-s}} a_{2^{n-s+1}-j}^{(s)} \sin \frac{(2k-1)\pi j}{2^{n-s+1}}, \quad k = 1, 2, \ldots, 2^{n-s} \tag{31}$$

pour $s = 1, 2, \ldots, n-1$.

Bref, le problème initial (22) se ramène au calcul du $(n-1)$-ième groupe de sommes (31). La transformation nécessaire à cette fin des coefficients $a_j^{(0)}$ est décrite par les formules (30).

A la seconde phase de l'algorithme il faut transformer les sommes (31) qui, après substitution de chaque s fixé,

$$z_k^{(0)}(1) = y_{2^{s-1}(2k-1)}, \quad k = 1, 2, \ldots, 2^{n-s},$$
$$b_j^{(0)}(1) = a^{(s)}_{2^{n-s+1}-j}, \quad j = 1, 2, \ldots, 2^{n-s},$$
$$l = n - s, \quad s = 1, 2, \ldots, n-1,$$

s'écrivent sous la forme suivante:

$$z_k^{(0)}(1) = \sum_{j=1}^{2^l} b_j^{(0)}(1) \sin\frac{(2k-1)\pi j}{2^{l+1}}, \quad k = 1, 2, \ldots, 2^l, \quad (32)$$

où $l = 1, 2, \ldots, n-1$. Les coefficients $b_j^{(0)}(1)$ et les fonctions $z_k^{(0)}(1)$ dépendent ici de l'indice l, or, comme le calcul de la somme (32) sera décrit pour un l fixé, on négligera cet indice.

Abordons la transformation de la somme (32). Représentons-la sous forme de deux termes en séparant les termes à indices pairs de j de ceux à indices impairs de j:

$$z_k^{(0)}(1) = \sum_{j=1}^{2^{l-1}} b_{2j}^{(0)}(1) \sin\frac{(2k-1)\pi j}{2^l} + \\ + \sum_{j=1}^{2^{l-1}} b_{2j-1}^{(0)}(1) \sin\frac{(2k-1)\pi(2j-1)}{2^{l+1}}. \quad (33)$$

En utilisant l'égalité

$$\sin\frac{(2k-1)(2j-2)\pi}{2^{l+1}} + \sin\frac{(2k-1)2j\pi}{2^{l+1}} = \\ = 2\cos\frac{(2k-1)\pi}{2^{l+1}} \sin\frac{(2k-1)(2j-1)\pi}{2^{l+1}},$$

écrivons le second terme de l'addition sous forme de deux sommes:

$$\sum_{j=1}^{2^{l-1}} b_{2j-1}^{(0)}(1) \sin\frac{\pi(2k-1)(2j-1)}{2^{l+1}} = \frac{1}{2\cos\frac{(2k-1)\pi}{2^{l+1}}} \times \\ \times \left[\sum_{j=1}^{2^{l-1}} b_{2j-1}^{(0)}(1) \sin\frac{(2k-1)\pi j}{2^l} + \sum_{j=1}^{2^{l-1}} b_{2j-1}^{(0)}(1) \sin\frac{(2k-1)\pi(j-1)}{2^l}\right] = \\ = \frac{1}{2\cos\frac{(2k-1)\pi}{2^{l+1}}} \left(b_{2^l-1}^{(0)}(1) \sin\frac{(2k-1)\pi}{2} + \sum_{j=1}^{2^{l-1}-1} (b_{2j+1}^{(0)}(1) + \right. \\ \left. + b_{2j-1}^{(0)}(1)) \sin\frac{(2k-1)\pi j}{2^l}\right). \quad (34)$$

Ajoutons que dans la seconde somme mise entre crochets on a substitué à l'indice j l'indice $j' + 1$.

Posons

$$b_j^{(1)}(1) = b_{2j-1}^{(0)} + b_{2j+1}^{(0)}(1), \quad j = 1, 2, \ldots, 2^{l-1} - 1,$$

$$b_{2^{l-1}}^{(1)}(1) = b_{2^l-1}^{(0)}(1),$$

$$b_j^{(1)}(2) = b_{2j}^{(0)}(1), \quad j = 1, 2, \ldots, 2^{l-1}$$

et portons (34) dans (33). On obtient l'expression

$$z_k^{(0)}(1) = \sum_{j=1}^{2^{l-1}} b_j^{(1)}(2) \sin \frac{(2k-1)\pi j}{2^l} + \\ + \frac{1}{2\cos\frac{(2k-1)\pi}{2^{l+1}}} \sum_{j=1}^{2^{l-1}} b_j^{(1)}(1) \sin \frac{(2k-1)\pi j}{2^l},$$

qui se vérifie pour $k = 1, 2, \ldots, 2^l$. En y portant au lieu de k $2^l - k + 1$, il vient

$$z_{2^l-k+1}^{(0)}(1) = -\sum_{j=1}^{2^{l-1}} b_j^{(1)}(2) \sin \frac{(2k-1)\pi j}{2^l} + \\ + \frac{1}{2\cos\frac{(2k-1)\pi}{2^{l+1}}} \sum_{j=1}^{2^{l-1}} b_j^{(1)}(1) \sin \frac{(2k-1)\pi j}{2^l}.$$

Par conséquent, si l'on pose

$$z_k^{(1)}(s) = \sum_{j=1}^{2^{l-1}} b_j^{(1)}(s) \sin \frac{(2k-1)\pi j}{2^l},$$

$$k = 1, 2, \ldots, 2^{l-1}, \quad s = 1, 2,$$

la somme initiale $z_k^{(0)}(1)$ peut alors être calculée suivant les formules

$$z_k^{(0)}(1) = z_k^{(1)}(2) + \frac{1}{2\cos\frac{(2k-1)\pi}{2^{l+1}}} z_k^{(1)}(1),$$

$$z_{2^l-k+1}^{(0)}(1) = -z_k^{(1)}(2) + \frac{1}{2\cos\frac{(2k-1)\pi}{2^{l+1}}} z_k^{(1)}(1),$$

$$k = 1, 2, \ldots, 2^{l-1}.$$

En résumé, le premier pas a abouti à l'apparition des sommes $z_k^{(1)}(1)$ et $z_k^{(1)}(2)$ dont chacune comprend deux fois moins de termes que la somme initiale $z_k^{(0)}(1)$ tout en possédant la même structure que $z_k^{(0)}(1)$. Par suite, le procédé de transformation, décrit plus haut, de la somme initiale peut s'appliquer séparément aux sommes $z_k^{(1)}(1)$ et $z_k^{(1)}(2)$. On obtient finalement les sommes $z_k^{(2)}(s)$, $s=1, 2, 3, 4$, qui conservent la structure de la somme initiale. En continuant la transformation, on obtient au m-ième pas les sommes

$$z_k^{(m)}(s)=\sum_{j=1}^{2^{l-m}} b_j^{(m)}(s)\sin\frac{(2k-1)\pi j}{2^{l-m+1}}, \tag{35}$$
$$k=1,\ 2,\ \ldots,\ 2^{l-m},\quad s=1,\ 2,\ \ldots,\ 2^m$$

pour chaque $m=0,\ 1,\ \ldots,\ l$, les coefficients $b_j^{(m)}(s)$ se déterminant par récurrence pour $s=1,\ 2,\ \ldots,\ 2^{m-1}$ suivant les formules

$$\begin{gathered}
b_j^{(m)}(2s-1)=b_{2j-1}^{(m-1)}(s)+b_{2j+1}^{(m-1)}(s),\\
j=1,\ 2,\ \ldots,\ 2^{l-m}-1,\quad m=1,\ 2,\ \ldots,\ l-1,\\
b_{2^{l-m}}^{(m)}(2s-1)=b_{2^{l-m+1}-1}^{(m-1)}(s),\quad m=1,\ 2,\ \ldots,\ l,\\
b_j^{(m)}(2s)=b_{2j}^{(m-1)}(s),\quad j=1,\ 2,\ \ldots,\ 2^{l-m},\\
m=1,\ 2,\ \ldots,\ l.
\end{gathered} \tag{36}$$

En outre, les sommes du m-ième pas sont liées aux sommes obtenues au $(m-1)$-ième pas au moyen des formules suivantes:

$$\begin{gathered}
z_k^{(m-1)}(s)=z_k^{(m)}(2s)+\frac{1}{2\cos\frac{\pi(2k-1)}{2^{l-m+2}}}z_k^{(m)}(2s-1),\\
z_{2^{l-m+1}-k+1}^{(m-1)}(s)=-z_k^{(m)}(2s)+\frac{1}{2\cos\frac{\pi(2k-1)}{2^{l-m+2}}}z_k^{(m)}(2s-1),
\end{gathered} \tag{37}$$
$$k=1,\ 2,\ \ldots,\ 2^{l-m},\quad s=1,\ 2,\ \ldots,\ 2^{m-1},\quad m=1,\ 2,\ \ldots,\ l.$$

En posant dans (35) $m=l$, il vient

$$z_1^{(l)}(s)=b_1^{(l)}(s),\quad s=1,\ 2,\ \ldots,\ 2^l. \tag{38}$$

Bref, les sommes $z^{(0)}(1)$ sont calculées de la façon suivante. En partant des coefficients donnés $b_j^{(0)}(1)$, $1\leqslant j\leqslant 2^l$, on calcule finalement à l'aide des formules (36) les coefficients $b_1^{(l)}(s)$, $1\leqslant s\leqslant 2^l$. Ces derniers sont ensuite utilisés en vertu de (38) en guise de données de base des relations de récurrence (37). En posant successivement dans (37) $m=l,\ l-1,\ \ldots,\ 1$, on obtient finalement $z_k^{(0)}(1)$ et, par suite, $y_{2^{s-1}(2k-1)}$.

Donc l'algorithme de calcul des sommes (22) est décrit par les formules (30), (36), (38).

R e m a r q u e. Dans les relations de récurrence (37) on peut éviter la division par $2\cos\frac{(2k-1)\pi}{2^{l-m+2}}$ avec la substitution

$$z_k^{(m)}(s)=\sin\frac{(2k-1)\pi}{2^{l-m+1}}\,w_k^{(m)}(s).$$

Les formules pour le calcul de $w_k^{(m)}(s)$ prennent dans ce cas la forme

$$\begin{aligned} w_k^{(m-1)}(s)&=2\cos\frac{\pi(2k-1)}{2^{l-m+2}}\,w_k^{(m)}(2s)+w_k^{(m)}(2s-1),\\ w_{2^{l-m+1}-k+1}^{(m-1)}(s)&=-2\cos\frac{\pi(2k-1)}{2^{l-m+2}}\,w_k^{(m)}(2s)+w_k^{(m)}(2s-1), \end{aligned} \tag{39}$$

$$k=1, 2, \ldots, 2^{l-m},\quad s=1, 2, \ldots, 2^{m-1},\quad m=l,\ l-1, \ldots, 1,$$

avec $w_1^{(l)}(s)=b_1^{(l)}(s)$, $s=1, 2, \ldots, 2^l$ et

$$z_k^{(0)}(1)=\sin\frac{(2k-1)\pi}{2^{l+1}}\,w_k^{(0)}(1),\quad k=1, 2, \ldots, 2^l. \tag{40}$$

Calculons maintenant le nombre d'opérations arithmétiques que coûte chacune des mises en œuvre de l'algorithme (30), (36)-(38). Admettons que les valeurs des fonctions trigonométriques sont connues d'avance.

Le calcul élémentaire fournit:

1) la mise en œuvre de (30) exige

$$Q_1=\sum_{p=1}^{n-1}2\,(2^{n-p}-1)=2\cdot 2^n-2\,(n+1)$$

opérations d'addition et de soustraction;

2) pour un l fixé la mise en œuvre de (36) exige

$$\bar{q}_l=\sum_{m=1}^{l-1}(2^{l-m}-1)\cdot 2^{m-1}=(l-2)\,2^{l-1}+1$$

opérations d'addition, et la mise en œuvre de (37)

$$\bar{\bar{q}}_l=\sum_{m=1}^{l}2\cdot 2^{l-m}\cdot 2^{m-1}=2l\cdot 2^{l-1}$$

opérations d'addition et

$$q_l^*=\sum_{m=1}^{l}2^{l-m}\cdot 2^{m-1}=l\cdot 2^{l-1} \tag{41}$$

opérations de multiplication. Finalement les formules (36) et (37) coûtent pour un l fixé

$$q_l=\bar{q}_l+\bar{\bar{q}}_l=(3l-2)\cdot 2^{l-1}+1 \tag{42}$$

opérations d'addition et q_l^* multiplications. Pour tous les $l = 1, 2, \ldots, n-1$ le coût revient à

$$Q_2 = \sum_{l=1}^{n-1} q_l = \sum_{l=1}^{n-1} [(3l-2)\cdot 2^{l-1}+1] = \frac{3}{2} n2^n - 4\cdot 2^n + n + 4$$

additions et

$$Q_3 = \sum_{l=1}^{n-1} q_l^* = \sum_{l=1}^{n-1} l2^{l-1} = \frac{n}{2} 2^n - 2^n + 1$$

multiplications.

Les algorithmes (30), (36)-(38) se caractérisent par les estimations suivantes du nombre d'opérations arithmétiques: $Q_+ = Q_1 + Q_2 = (3n/2 - 2)\, 2^n - n + 2$ additions et $Q_* = (n/2-1)\, 2^n + 1$ multiplications. Si on néglige la différence entre les opérations d'addition et de multiplication, leur nombre total s'élève à

$$Q = Q_1 + Q_2 + Q_3 = (2 \log_2 N - 3)\, N - \log_2 N + 3,\ N = 2^n.$$

A titre de comparaison, donnons l'estimation du nombre d'opérations exigé avec le calcul de toutes les sommes (22) par sommation directe. On aura $(2^n-1)^2$ multiplications et $(2^n-2)\,(2^n-1)$ additions, en tout $\widetilde{Q} = (N-1)\,(2N-3)$ opérations. Par exemple, pour $N = 128$ $(n = 7)$, on aura $Q = 1404$ opérations (dont 321 multiplications) pour la mise en œuvre de l'algorithme et $\widetilde{Q} = 32\,131$ opérations (dont 15 873 multiplications) pour la mise en œuvre de l'algorithme par sommation directe.

Notons que l'utilisation dans l'algorithme au lieu de (37) et (38) des relations (39) et (40) aboutit aux estimations suivantes du nombre d'opérations: $Q_+ = \left(\frac{3}{2}n-2\right)2^n - n + 2$ additions et $Q_* = \frac{n}{2}\, 2^n - 1$ multiplications, en tout $Q = (2\log_2 N - 2)\, N - \log_2 N + 1$, $N = 2^n$; ce nombre est quelque peu supérieur à celui de l'algorithme (30), (36)-(38).

Bref, le problème 1 posé plus haut est résolu. Passons maintenant au problème 2 sur le développement en sinus rapprochés. En posant $N = 2^n$, écrivons la somme figurant dans le problème 2 sous la forme suivante:

$$y_k = \sum_{j=1}^{2^n} a_j \sin \frac{(2k-1)\,\pi j}{2^{n+1}}, \quad k = 1, 2, \ldots, 2^n. \tag{43}$$

En confrontant (43) et (32), on trouve que le calcul des sommes (43) suivant les sinus rapprochés constitue la seconde phase du calcul des sommes (22) de l'algorithme exposé plus haut, si dans (32) on

pose $l = n$. Par conséquent, si l'on pose

$$z_k^{(0)}(1) = y_k, \quad k = 1, 2, \ldots, 2^n,$$
$$b_j^{(0)}(1) = a_j, \quad j = 1, 2, \ldots, 2^n,$$

les formules (36)-(38) pour $l = n$ décrivent l'algorithme de calcul des sommes (43). En posant dans les formules (41) et (42) $l = n$, on obtient les estimations suivantes du coût de la mise en œuvre de l'algorithme: $Q_+ = q_n = \left(\frac{3}{2}n - 1\right)2^n + 1$ opérations d'addition et $Q_* = q_n^* = \frac{n}{2}2^n$ opérations de multiplication, en tout $Q = (2\log_2 N - 1)N + 1$, $N = 2^n$. Les sommes (43) se calculent donc en, à peu près, le même nombre d'opérations arithmétiques que les sommes (22).

Rappelons que les sommes (43) sont utilisées pour le calcul des coefficients de Fourier de la fonction de maille a_i donnée pour $i = 1, 2, \ldots, N$. Pour le rétablissement de la fonction sur la base des coefficients de Fourier donnés il faut calculer les sommes

$$y_j = \sum_{k=1}^{2^n} a_k \sin\frac{(2k-1)\pi j}{2^{n+1}}, \quad j = 1, 2, \ldots, 2^n. \qquad (43')$$

En utilisant pour $j \neq 2^n$ la relation

$$\sin\frac{(2k-1)\pi j}{2^{n+1}} = \frac{1}{2\cos\frac{\pi j}{2^{n+1}}}\left[\sin\frac{(k-1)\pi j}{2^n} + \sin\frac{k\pi j}{2^n}\right],$$

il vient

$$y_j = \frac{1}{2\cos\frac{\pi j}{2^{n+1}}}\left[\sum_{k=1}^{2n} a_k \sin\frac{(k-1)\pi j}{2^n} + \sum_{k=1}^{2^n} a_k \sin\frac{k\pi j}{2^n}\right] =$$
$$= \frac{1}{2\cos\frac{\pi j}{2^{n+1}}}\sum_{k=1}^{2^n-1} a_k^{(0)} \sin\frac{k\pi j}{2^n}, \quad j = 1, 2, \ldots, 2^{n-1},$$

où $a_k^{(0)}$ se calculent suivant la formule $a_k^{(0)} = a_k + a_{k+1}$, $k = 1, 2, \ldots, 2^n - 1$. En comparant la somme obtenue avec (22), on trouve que le problème s'est réduit à la résolution du problème 1.

Pour le calcul de y_{2^n}, on obtient la formule

$$y_{2^n} = \sum_{k=1}^{2^n} a_k(-1)^{k-1} = \sum_{k=1}^{2^{n-1}} (a_{2k-1} - a_{2k}).$$

La sommation est ici effectuée directement.

En ce qui concerne le nombre d'opérations que coûte l'algorithme décrit, l'estimation $Q = 2N\log_2 N - \log_2 N$ se vérifie.

3. Développement en cosinus. Examinons maintenant l'algorithme de résolution du problème 3 comprenant des calculs des sommes (13) pour $N = 2^n$. On a

$$y_k = \sum_{j=0}^{2^n} a_j^{(0)} \cos \frac{k\pi j}{2^n}, \quad k = 0, 1, \ldots, 2^n, \tag{44}$$

où on a introduit la notation $a_j^{(0)} = a_j$.

Le principe de construction de l'algorithme est exactement le même que pour le cas de développement en sinus et comprend deux étapes. Dans la première étape on rassemble les termes de la somme aux indices j et $2^n - j$ pour $j = 0, 1, \ldots, 2^{n-1} - 1$, ensuite aux indices j et $2^{n-1} - j$, $j = 0, 1, \ldots, 2^{n-2} - 1$, etc.

Au bout du p-ième pas on aura

$$\begin{gathered} y_{2^{s-1}(2k-1)} = \sum_{j=0}^{2^{n-s}-1} a^{(s)}_{2^{n-s+1}-j} \cos \frac{(2k-1)\pi j}{2^{n-s+1}}, \\ k = 1, 2, \ldots, 2^{n-s}, \quad s = 1, 2, \ldots, p, \\ y_{2^p k} = \sum_{j=0}^{2^{n-p}} a_j^{(p)} \cos \frac{k\pi j}{2^{n-p}}, \quad k = 0, 1, \ldots, 2^{n-p}. \end{gathered} \tag{45}$$

Ces formules se vérifient pour $p = 1, 2, \ldots, n$. Les coefficients $a_j^{(p)}$ se déterminent par récurrence

$$\begin{aligned} a_j^{(p)} &= a_j^{(p-1)} + a^{(p-1)}_{2^{n-p+1}-j}, \\ a^{(p)}_{2^{n-p+1}-j} &= a_j^{(p-1)} - a^{(p-1)}_{2^{n-p+1}-j}, \quad j = 0, 1, \ldots, 2^{n-p} - 1, \\ a^{(p)}_{2^{n-p}} &= a^{(p-1)}_{2^{n-p}}, \quad p = 1, 2, \ldots, n. \end{aligned} \tag{46}$$

En posant dans (45) $s = p = n$, il vient

$$y_0 = a_0^{(n)} + a_1^{(n)}, \quad y_{2^n} = a_0^{(n)} - a_1^{(n)}, \quad y_{2^{n-1}} = a_2^{(n)}, \tag{47}$$

les y_k restants se déterminent suivant les formules

$$\begin{gathered} y_{2^{s-1}(2k-1)} = \sum_{j=0}^{2^{n-s}-1} a^{(s)}_{2^{n-s+1}-j} \cos \frac{(2k-1)\pi j}{2^{n-s+1}}, \\ k = 1, 2, \ldots, 2^{n-s}, \quad s = 1, 2, \ldots, n-1. \end{gathered}$$

Pour chaque s fixé les substitutions

$$\begin{gathered} z_k^{(0)}(1) = y_{2^{s-1}(2k-1)}, \quad k = 1, 2, \ldots, 2^{n-s}, \\ b_j^{(0)}(1) = a^{(s)}_{2^{n-s+1}-j}, \quad j = 0, 1, \ldots, 2^{n-s} - 1, \\ l = n - s, \quad s = 1, 2, \ldots, n-1 \end{gathered}$$

aboutissent au calcul des sommes suivantes:

$$z_k^{(0)}(1)=\sum_{j=0}^{2^l-1} b_j^{(0)}(1)\cos\frac{(2k-1)\pi j}{2^{l+1}},\quad k=1, 2, \ldots, 2^l,$$
$$l=1, 2, \ldots, n-1. \tag{48}$$

Dans la seconde étape de la mise en œuvre de l'algorithme on effectue le calcul des sommes (48). Comme auparavant, on sépare successivement les termes à indices j pairs de ceux à indices j impairs et on aboutit aux relations de récurrence suivantes:

$$\begin{gathered} z_k^{(m-1)}(s)=z_k^{(m)}(2s)+\frac{1}{2\cos\frac{(2k-1)\pi}{2^{l-m+2}}}z_k^{(m)}(2s-1), \\ z_{2^{l-m+1}-k+1}^{(m)}(s)=z_k^{(m)}(2s)-\frac{1}{2\cos\frac{(2k-1)\pi}{2^{l-m+2}}}z_k^{(m)}(2s-1), \\ k=1, 2, \ldots, 2^{l-m},\quad s=1, 2, \ldots, 2^{m-1},\quad m=1, 2, \ldots, l \end{gathered} \tag{49}$$

pour le calcul de

$$z_k^{(m)}(s)=\sum_{j=0}^{2^{l-m}-1} b_j^{(m)}(s)\cos\frac{(2k-1)\pi j}{2^{l-m+1}}, \tag{50}$$
$$k=1, 2, \ldots, 2^{l-m},\quad s=1, 2, \ldots, 2^m$$

pour $m=0, 1, \ldots, l$. Les coefficients $b_j^{(m)}(s)$ se déterminent également par récurrence pour $s=1, 2, \ldots, 2^{m-1}$, en partant de $b_j^{(0)}(1)$ à l'aide des formules

$$\begin{gathered} b_j^{(m)}(2s-1)=b_{2j-1}^{(m-1)}(s)+b_{2j+1}^{(m-1)}(s), \\ j=1, 2, \ldots, 2^{l-m}-1,\quad m=1, 2, \ldots, l-1, \\ b_0^{(m)}(2s-1)=b_1^{(m-1)}(s),\quad m=1, 2, \ldots, l, \\ b_j^{(m)}(2s)=b_{2j}^{(m-1)}(s), \\ j=0, 1, \ldots, 2^{l-m}-1,\quad m=1, 2, \ldots, l. \end{gathered} \tag{51}$$

En posant dans (50) $m=l$, on obtient les données de base pour les relations (49)

$$z_1^{(l)}(s)=b_0^{(l)}(s),\quad s=1, 2, \ldots, 2^l. \tag{52}$$

Ainsi donc, les algorithmes de calcul des sommes (44) se décrivent au moyen des formules (46), (47), (49), (51) et (52).

Un calcul élémentaire du nombre d'opérations arithmétiques que coûte la mise en œuvre de l'algorithme donne: $Q_+ =$

$= (3/2n - 2)\, 2^n + n + 2$ opérations d'addition et $Q_* =$
$= (n/2 - 1)\, 2^n + 1$ opérations de multiplication, en tout

$$Q = Q_+ + Q_* = (2 \log_2 N - 3)\, N + \log_2 N + 3, \quad N = 2^n.$$

Notons que, comme dans l'algorithme précédent, on peut ici procéder dans les relations (49) à la substitution

$$z_k^{(m)}(s) = \sin \frac{(2k-1)\pi}{2^{l-m+1}}\, w_k^{(m)}(s);$$

en outre, il s'ensuit de (52) que $w_1^{(l)}(s) = b_0^{(l)}(s)$, $s = 1, 2, \ldots, 2^l$.

Les formules de récurrence pour $w_k^{(m)}(s)$ prennent la forme

$$w_k^{(m-1)}(s) = 2\cos \frac{(2k-1)\pi}{2^{l-m+2}}\, w_k^{(m)}(2s) + \bar{w}_k^{(m)}(2s-1),$$

$$w_{2^{l-m+1}-k+1}^{(m-1)}(s) = 2\cos \frac{(2k-1)\pi}{2^{l-m+2}}\, w_k^{(m)}(s) - w_k^{(m)}(2s-1),$$

$$k = 1, 2, \ldots, 2^{l-m}, \quad s = 1, 2, \ldots, 2^{m-1}, \quad m = 1, 2, \ldots, l.$$

4. Transformation d'une fonction de maille périodique réelle. Le problème 4 sur la transformation d'une fonction de maille périodique réelle consiste à rétablir la fonction au moyen des formules (17), les coefficients de Fourier a_j et $\bar{a}_j$ étant donnés, et à trouver les coefficients de la fonction considérée suivant les formules (18).

Soient $N = 2^n$ et les coefficients de Fourier connus. Il faut alors calculer les sommes

$$y_k = \sum_{j=0}^{2^{n-1}} a_j^{(0)} \cos \frac{2k\pi j}{2^n} + \sum_{j=1}^{2^{n-1}-1} \bar{a}_j^{(0)} \sin \frac{2k\pi j}{2^n}, \qquad k = 0, 1, \ldots, 2^n - 1. \tag{53}$$

Construisons l'algorithme correspondant. A cette fin substituons dans (53) $2^n - k$ à k. Compte tenu des égalités

$$\cos \frac{2(2^n-k)\pi j}{2^n} = \cos \frac{2k\pi j}{2^n}, \quad \sin \frac{2(2^n-k)\pi j}{2^n} = -\sin \frac{2k\pi j}{2^n},$$

on obtient que y_k peut être calculé suivant les formules

$$\begin{gathered} y_k = \bar{y}_k + \bar{\bar{y}}_k, \\ y_{2^n-k} = \bar{y}_k - \bar{\bar{y}}_k, \quad k = 1, 2, \ldots, 2^{n-1} - 1, \\ y_0 = \bar{y}_0, \quad y_{2^{n-1}} = \bar{y}_{2^{n-1}}, \end{gathered} \tag{54}$$

où

$$\bar{y}_k = \sum_{j=0}^{2^{n-1}}{}' a_j^{(0)} \cos \frac{k\pi j}{2^{n-1}}, \quad k = 0, 1, \ldots, 2^{n-1}, \tag{55}$$

$$\bar{\bar{y}}_k = \sum_{j=1}^{2^{n-1}-1} \bar{a}_j^{(0)} \sin \frac{k\pi j}{2^{n-1}}, \quad k = 1, 2, \ldots, 2^{n-1}-1. \tag{56}$$

En résumé, le calcul des sommes (53) se réduit à celui des sommes (55) et (56) et à l'utilisation subséquente des formules (54).

En comparant les formules (55) et (56) avec les formules (44) et (22) on observe que les sommes (55) et (56) peuvent être calculées suivant les algorithmes des points 2 et 3 en y substituant $n - 1$ à n.

Evaluons maintenant le nombre d'opérations arithmétiques que coûte le calcul des sommes (53) par la méthode indiquée. A partir des estimations du nombre d'opérations obtenu pour l'algorithme du point 2 on déduit que les sommes (56) s'obtiennent en $Q_+ = (3n/4 - 7/4)\, 2^n - n + 3$ opérations d'addition et en $Q_* = (n/4 - 3/4)\, 2^n + 1$ opérations de multiplication. Les estimations de l'algorithme au point 3 fournissent pour les sommes (55) les valeurs suivantes: $Q_+ = (3n/4 - 7/4)\, 2^n + n + 1$ additions et $Q_* = (n/4 - 3/4)\, 2^n + 1$ multiplications. En y adjoignant $Q_+ = 2^n - 2$ additions pour la mise en œuvre de (54), on obtient pour l'algorithme construit $Q_+ = (3n/2 - 5/2)\, 2^n + 2$ additions et $Q_* = (n/2 - 3/2)\, 2^n + 2$ multiplications, en tout $Q = (2 \log_2 N - 4) N + 4$, $N = 2^n$.

Abordons maintenant le calcul des coefficients de Fourier de la fonction de maille périodique réelle. Le problème réside en l'obtention des sommes

$$y_k = \sum_{j=0}^{2^n-1} a_j^{(0)} \cos \frac{2k\pi j}{2^n}, \quad k = 0, 1, \ldots, 2^{n-1}, \tag{57}$$

$$\bar{y}_k = \sum_{j=1}^{2^n-1} a_j^{(0)} \sin \frac{2k\pi j}{2^n}, \quad k = 1, 2, \ldots, 2^{n-1}-1, \tag{58}$$

où $a_j^{(0)}$ est la fonction donnée.

L'algorithme de calcul de (57) et (58) est apparenté à ceux des points 2 et 3, tout en différant par certains détails. Ici, dans la première phase sont d'abord rassemblés les termes des sommes (57) et (58) aux indices j et $2^{n-1} + j$ pour $j = 0, 1, \ldots, 2^{n-1} - 1$, ensuite, ceux aux indices j et $2^{n-2} + j$ pour $j = 0, 1, \ldots, 2^{n-2} - 1$, etc. Décrivons en plus de détails le premier pas du procédé de ras-

semblement successif des termes d'addition sur l'exemple de la somme (57). La transformation de la somme (58) s'effectue de façon analogue.

Représentons donc (57) sous la forme suivante:

$$y_k = \sum_{j=0}^{2^{n-1}-1} a_j^{(0)} \cos \frac{2k\pi j}{2^n} + \sum_{j=2^{n-1}}^{2^n-1} a_j^{(0)} \cos \frac{2k\pi j}{2^n}$$

et effectuons dans la seconde somme la substitution en posant $j = 2^{n-1} + j'$. Cela donne

$$y_k = \sum_{j=0}^{2^{n-1}-1} [a_j^{(0)} + (-1)^k a_{2^{n-1}+j}^{(0)}] \cos \frac{2k\pi j}{2^n}, \quad k = 0, 1, \ldots, 2^{n-1}.$$

En posant

$$\begin{aligned} a_j^{(1)} &= a_j^{(0)} + a_{2^{n-1}+j}^{(0)}, \\ a_{2^{n-1}+j}^{(1)} &= a_j^{(0)} - a_{2^{n-1}+j}^{(0)}, \quad j = 0, 1, \ldots, 2^{n-1}-1, \end{aligned} \tag{59}$$

on obtient au lieu de (57) les sommes suivantes:

$$\begin{aligned} y_{2k-1} &= \sum_{j=0}^{2^{n-1}-1} a_{2^{n-1}+j}^{(1)} \cos \frac{(2k-1)\pi j}{2^{n-1}}, \quad k = 1, 2, \ldots, 2^{n-2}, \\ y_{2k} &= \sum_{j=0}^{2^{n-1}-1} a_j^{(1)} \cos \frac{2k\pi j}{2^{n-1}}, \quad k = 0, 1, \ldots, 2^{n-2}. \end{aligned} \tag{60}$$

De façon analogue, au lieu de (58) on obtient les sommes

$$\begin{aligned} \bar{y}_{2k-1} &= \sum_{j=1}^{2^{n-1}-1} a_{2^{n-1}+j}^{(1)} \sin \frac{(2k-1)\pi j}{2^{n-1}}, \quad k = 1, 2, \ldots, 2^{n-2}, \\ \bar{y}_{2k} &= \sum_{j=1}^{2^{n-1}-1} a_j^{(1)} \text{ in} \frac{2k\pi j}{2^{n-1}}, \quad k = 1, 2, \ldots, 2^{n-2}-1, \end{aligned} \tag{61}$$

où $a_j^{(1)}$ sont définis dans (59). Sur ce, le premier pas s'achève. Au cours du second pas on transforme les sommes (60) et (61) suivant le procédé décrit. Finalement au bout du p-ième pas on a

$$y_{2^{s-1}(2k-1)} = \sum_{j=0}^{2^{n-s}-1} a_{2^{n-s}+j}^{(s)} \cos \frac{(2k-1)\pi j}{2^{n-s}},$$

$$k = 1, 2, \ldots, 2^{n-s-1}, \quad s = 1, 2, \ldots, p, \tag{62}$$

$$y_{2^pk} = \sum_{j=0}^{2^{n-p}-1} a_j^{(p)} \cos \frac{2k\pi j}{2^{n-p}}, \quad k=0, 1, \ldots, 2^{n-p-1},$$

où $p=1, 2, \ldots, n-1$ et

$$\overline{y}_{2^{s-1}(2k-1)} = \sum_{j=1}^{2^{n-s}-1} a_{2^{n-s}+j}^{(s)} \sin \frac{(2k-1)\pi j}{2^{n-s}},$$

$$k=1, 2, \ldots, 2^{n-s-1}, \quad s=1, 2, \ldots, p, \tag{63}$$

$$\overline{y}_{2^pk} = \sum_{j=1}^{2^{n-p}-1} a_j^{(p)} \sin \frac{2k\pi j}{2^{n-p}}, \quad k=1, 2, \ldots, 2^{n-p-1}-1,$$

où $p=1, 2, \ldots, n-2$. Les coefficients $a_j^{(p)}$ s'obtiennent par récurrence suivant les formules

$$\begin{aligned} a_j^{(p)} &= a_j^{(p-1)} + a_{2^{n-p}+j}^{(p-1)}, \quad j=0, 1, \ldots, 2^{n-p}-1, \\ a_{2^{n-p}+j}^{(p)} &= a_j^{(p-1)} - a_{2^{n-p}+j}^{(p-1)}, \quad p=1, 2, \ldots, n. \end{aligned} \tag{64}$$

En posant dans (62) $p=n$ et $s=p=n-1$, il vient

$$\begin{aligned} y_0 &= a_0^{(n-1)} + a_1^{(n-1)} = a_0^{(n)}, \\ y_{2^{n-1}} &= a_0^{(n-1)} - a_1^{(n-1)} = a_1^{(n)}, \\ y_{2^{n-2}} &= a_2^{(n-1)}, \end{aligned} \tag{65}$$

tandis qu'à partir de (63) pour $p=n-2$ on trouve

$$\overline{y}_{2^{n-2}} = a_1^{(n-2)} - a_3^{(n-2)} = a_3^{(n-1)}. \tag{66}$$

Les y_k et $\overline{y}_k$ restants s'obtiennent suivant les formules

$$y_{2^{s-1}(2k-1)} = \sum_{j=0}^{2^{n-s}-1} a_{2^{n-s}+j}^{(s)} \cos \frac{(2k-1)\pi j}{2^{n-s}},$$

$$\overline{y}_{2^{s-1}(2k-1)} = \sum_{j=1}^{2^{n-s}-1} a_{2^{n-s}+j}^{(s)} \sin \frac{[(2k-1)\pi j}{2^{n-s}},$$

$$k=1, 2, \ldots, 2^{n-s-1}, \quad s=1, 2, \ldots, n-2.$$

Procédons aux substitutions pour un s fixé:

$$z_k^{(0)}(1) = y_{2^{s-1}(2k-1)}, \quad \overline{z}_k^{(0)}(1) = \overline{y}_{2^{s-1}(2k-1)},$$

$$k=1, 2, \ldots, 2^{n-s-1}, \quad b_j^{(0)}(1) = a_{2^{n-s}+j}^{(s)}, \quad j=0, 1, \ldots, 2^{n-s}-1,$$

$$l=n-s, \quad s=1, 2, \ldots, n-2.$$

On aboutit ainsi au calcul des sommes

$$\begin{aligned} z_k^{(0)}(1) &= \sum_{j=0}^{2^l-1} b_j^{(0)}(1)\cos\frac{(2k-1)\pi j}{2^l}, \\ \bar{z}_k^{(0)}(1) &= \sum_{j=1}^{2^l-1} b_j^{(0)}(1)\sin\frac{(2k-1)\pi j}{2^l}, \end{aligned} \qquad (67)$$

$$k=1, 2, \ldots, 2^{l-1}, \quad l=2, 3, \ldots, n-1.$$

Au cours de la seconde phase de l'algorithme on calcule les sommes (67). Comme dans le cas de l'algorithme du point 2, on transforme ces sommes par séparation des termes aux indices j pairs des termes aux indices j impairs et on utilise les égalités

$$\sin\frac{(2k-1)(2j-2)\pi}{2^{l-m+1}} + \sin\frac{(2k-1)2j\pi}{2^{l-m+1}} = $$
$$= 2\cos\frac{(2k-1)\pi}{2^{l-m+1}}\sin\frac{(2k-1)(2j-1)\pi}{2^{l-m+1}},$$

$$\cos\frac{(2k-1)(2j-2)\pi}{2^{l-m+1}} + \cos\frac{(2k-1)2j\pi}{2^{l-m+1}} = $$
$$= 2\cos\frac{(2k-1)\pi}{2^{l-m+1}}\cos\frac{(2k-1)(2j-1)\pi}{2^{l-m+1}}$$

pour $m = 1, 2, \ldots$ On obtient ainsi les formules de récurrence suivantes:

$$\begin{aligned} z_k^{(m-1)}(s) &= z_k^{(m)}(2s) + \frac{1}{2\cos\frac{(2k-1)\pi}{2^{l-m+1}}} z_k^{(m)}(2s-1), \\ z_{2^{l-m}-k+1}^{(m-1)}(s) &= z_k^{(m)}(2s) - \frac{1}{2\cos\frac{(2k-1)\pi}{2^{l-m+1}}} z_k^{(m)}(2s-1), \\ \bar{z}_k^{(m-1)}(s) &= \bar{z}_k^{(m)}(2s) + \frac{1}{2\cos\frac{(2k-1)\pi}{2^{l-m+1}}} \bar{z}_k^{(m)}(2s-1), \\ \bar{z}_{2^{l-m}-k+1}^{(m-1)}(s) &= -\bar{z}_k^{(m)}(2s) + \frac{1}{2\cos\frac{(2k-1)\pi}{2^{l-m+1}}} \bar{z}_k^{(m)}(2s-1), \end{aligned} \qquad (68)$$

$$k=1, 2, \ldots, 2^{l-m-1}, \quad s=1, 2, \ldots, 2^{m-1}, \quad m=1, 2, \ldots, l-1$$

permettant de calculer de proche en proche les sommes

$$
\begin{aligned}
z_k^{(m)}(s) &= \sum_{j=0}^{2^{l-m}-1} b_j^{(m)}(s) \cos \frac{(2k-1)\pi j}{2^{l-m}}, \\
\bar{z}_k^{(m)}(s) &= \sum_{j=1}^{2^{l-m}-1} b_j^{(m)}(s) \sin \frac{(2k-1)\pi j}{2^{l-m}}, \\
k &= 1, 2, \ldots, 2^{l-m-1}, \quad s = 1, 2, \ldots, 2^m
\end{aligned}
\tag{69}
$$

pour $m = 0, 1, \ldots, l-1$.

Les coefficients $b_j^{(m)}(s)$ s'obtiennent également par récurrence pour $s = 1, 2, \ldots, 2^{m-1}$ en partant des $b_j^{(0)}(1)$ donnés suivant les formules

$$
\begin{aligned}
b_j^{(m)}(2s-1) &= b_{2j-1}^{(m-1)}(s) + b_{2j+1}^{(m-1)}(s), \quad j = 1, 2, \ldots, 2^{l-m}-1, \\
b_0^{(m)}(2s-1) &= b_1^{(m-1)}(s) - b_{2^{l-m+1}-1}^{(m-1)}(s), \\
b_j^{(m)}(2s) &= b_{2j}^{(m-1)}(s), \quad j = 0, 1, \ldots, 2^{l-m}-1, \\
s &= 1, 2, \ldots, 2^{m-1}, \quad m = 1, 2, \ldots, l-1.
\end{aligned}
\tag{70}
$$

En posant dans (69) $m = l-1$, on obtient les valeurs initiales pour les relations (68)

$$
z_1^{(l-1)}(s) = b_0^{(l-1)}(s), \quad \bar{z}_1^{(l-1)}(s) = b_1^{(l-1)}(s), \quad s = 1, 2, \ldots, 2^{l-1}. \tag{71}
$$

Bref, l'algorithme du calcul simultané des sommes (57) et (58) se décrit au moyen des formules (64)-(66), (68), (70) et (71). Notons que, comme dans les algorithmes des points 2 et 3, il est possible de procéder ici dans les relations (68) à des substitutions

$$
\begin{aligned}
z_k^{(m)}(s) &= \sin \frac{(2k-1)\pi}{2^{l-m}} w_k^{(m)}(s), \\
\bar{z}_k^{(m)}(s) &= \sin \frac{(2k-1)\pi}{2^{l-m}} \bar{w}_k^{(m)}(s),
\end{aligned}
$$

qui permettent d'échapper à la division par $2 \cos \frac{(2k-1)\pi}{2^{l-m+1}}$.

Un calcul élémentaire du nombre d'opérations arithmétiques que coûte la mise en œuvre de l'algorithme donne : $Q_+ = 3n/2 \cdot 2^n - 1$ additions et $Q_* = (n/2 - 3/2)\, 2^n + 2$ multiplications, en tout $Q = (2 \log_2 N - 3/2)\, N + 1$, $N = 2^n$.

Le calcul des coefficients de Fourier et le rétablissement de la fonction de maille périodique réelle suivant l'algorithme proposé exigent donc $O(N \ln N)$ opérations arithmétiques.

5. Transformation de la fonction de maille périodique complexe. Abordons maintenant le problème 5 concernant le calcul des coef-

ficients de Fourier et le rétablissement de la fonction de maille périodique complexe. Au point 1 on a montré que le problème se ramenait au calcul des sommes (21) qui, au cas où $N = 2^n$, prennent la forme

$$y_k = \sum_{j=0}^{2^n-1} a_j^{(0)} e^{\frac{2k\pi j}{2^n} i}, \quad k = 0, 1, \ldots, 2^n - 1, \tag{72}$$

où $a_j^{(0)}$ sont des nombres complexes.

L'algorithme de calcul des sommes (72) se construit de la même façon que celui de calcul des coefficients de Fourier de la fonction périodique réelle. Lors de la première phase, on rassemble les termes des sommes (72), d'abord aux indices j et $2^{n-1} + j$ pour $j = 0, 1, \ldots$ $\ldots, 2^{n-1} - 1$, puis aux indices j et $2^{n-2} + j$ pour $j = 0, 1, \ldots$ $\ldots, 2^{n-2} - 1$, etc. Compte tenu de l'égalité $e^{\pi k i} = (-1)^k$, on obtient finalement pour le p-ième pas les sommes suivantes:

$$y_{2^{s-1}(2k-1)} = \sum_{j=0}^{2^{n-s}-1} a^{(s)}_{2^{n-s}+j} e^{\frac{(2k-1)\pi j}{2^{n-s}} i},$$

$$k = 1, 2, \ldots, 2^{n-s}, \quad s = 1, 2, \ldots, p, \tag{73}$$

$$y_{2^p k} = \sum_{j=0}^{2^{n-p}-1} a_j^{(p)} e^{\frac{2k\pi j}{2^{n-s}} i}, \quad k = 0, 1, \ldots, 2^{n-p} - 1,$$

où les coefficients $a_j^{(p)}$ s'obtiennent par récurrence suivant les formules (64).

En posant dans (73) $s = p = n$, il vient

$$y_0 = a_0^{(n)}, \quad y_{2^{n-1}} = a_1^{(n)}, \tag{74}$$

tandis que les y_k restants s'obtiennent suivant les formules

$$y_{2^{s-1}(2k-1)} = \sum_{j=0}^{2^{n-s}-1} a^{(s)}_{2^{n-s}+j} e^{\frac{(2k-1)\pi j}{2^{n-s}} i},$$

$$k = 1, 2, \ldots, 2^{n-s}, \quad s = 1, 2, \ldots, n - 1.$$

Procédons maintenant pour un j fixé à des substitutions en posant

$$z_k^{(0)}(l) = y_{2^{s-1}(2k-1)}, \quad k = 1, 2, \ldots, 2^{n-s},$$

$$b_j^{(0)}(l) = a^{(s)}_{2^{n-s}+j}, \quad j = 0, 1, \ldots, 2^{n-s} - 1,$$

$$l = n - s, \quad s = 1, 2, \ldots, n - 1,$$

et passons au calcul des sommes

$$z_k^{(0)}(1)=\sum_{j=0}^{2^l-1} b_j^{(0)}(1)\, e^{\frac{(2k-1)\pi j}{2^l}i}, \quad k=1,\,2,\,\ldots,\,2^l \tag{75}$$

pour $l = 1, 2, \ldots, n-1$.

La seconde phase de l'algorithme consistant à calculer les sommes (75) se construit, comme auparavant, par séparation des termes aux indices j pairs et impairs compte tenu des égalités

$$e^{\frac{(2k-1)(2j-2)\pi}{2^{l-m+1}}}+e^{\frac{(2k-1)2j\pi}{2^{l-m+1}}}=2\cos\frac{(2k-1)\pi}{2^{l-m+1}}\, e^{\frac{(2k-1)(2j-1)\pi}{2^{l-m+1}}}.$$

On obtient les formules de récurrence

$$\begin{aligned} z_k^{(m-1)}(s)&=z_k^{(m)}(2s)+\frac{1}{2\cos\frac{(2k-1)\pi}{2^{l-m+1}}}\, z_k^{(m)}(2s-1),\\ z_{2^{l-m}+k}^{(m-1)}(s)&=z_k^{(m)}(2s)-\frac{1}{2\cos\frac{(2k-1)\pi}{2^{l-m+1}}}\, z_k^{(m)}(2s-1),\end{aligned} \tag{76}$$

$$k=1,\,2,\,\ldots,\,2^{l-m},\quad s=1,\,2,\,\ldots,\,2^{m-1},\quad m=1,\,2,\,\ldots,\,l-1$$

pour le calcul des sommes

$$z_k^{(m)}(s)=\sum_{j=0}^{2^{l-m}-1} b_j^{(m)}(s)\, e^{\frac{(2k-1)\pi j}{2^{l-m}}i}, \tag{77}$$

$$k=1,\,2,\,\ldots,\,2^{l-m},\quad s=1,\,2,\,\ldots,\,2^m$$

pour $m = 0, 1, \ldots, l-1$. Les coefficients $b_j^{(m)}$ se calculent suivant les formules de récurrence (70). Il ne reste qu'à indiquer les valeurs initiales de (76). En posant dans (77) $m = l-1$, il vient

$$\begin{aligned} z_1^{(l-1)}(s)&=b_0^{(l-1)}(s)+ib_1^{(l-1)}(s),\\ z_2^{(l-1)}(s)&=b_0^{(l-1)}(s)-ib_1^{(l-1)}(s),\quad s=1,\,2,\,\ldots,\,2^{l-1}.\end{aligned} \tag{78}$$

Bref, l'algorithme de calcul des sommes (72) est décrit par les formules (64), (70), (74), (76) et (78). Notons que l'algorithme construit ne contient pas (à l'exclusion de la très simple formule (78)) d'opérations de multiplication des nombres complexes. Aussi dans les formules fournies est-il aisé d'isoler les parties réelle et imaginaire des grandeurs calculées. Cela facilite la mise en œuvre de l'algorithme sur l'ordinateur démuni d'arithmétique complexe. Ensuite, il peut aussi s'avérer utile de procéder dans les relations (76) à la substitution

$$z_k^{(m)}(s)=\sin\frac{(2k-1)\pi}{2^{l-m}}\, w_k^{(m)}(s).$$

Evaluons maintenant le nombre d'opérations arithmétiques que coûte la construction de l'algorithme. On obtient $Q_+ = (3n/2 - 1/2)\, 2^n$ additions de nombres complexes et $Q_* = (n/2 - 3/2)\, 2^n$ multiplications de nombres complexes par un nombre réel. Si l'on exprime ces valeurs en termes d'opérations sur des nombres réels, on obtient $Q_+ = (3n - 1)\, 2^n$ additions sur des nombres réels et $Q_* = (n - 3)\, 2^n$ multiplications sur des nombres réels, en tout $Q = (4 \log_2 N - 4)\, N$, $N = 2^n$ opérations sur des nombres réels. Cette estimation dépasse deux fois celle obtenue au point 4 pour une fonction de maille périodique réelle, ce qui d'ailleurs est naturel vu que dans le cas complexe étudié on est obligé de traiter une quantité deux fois plus grande de nombres réels.

Sur ce, on achève l'étude des algorithmes de transformation discrète rapide de Fourier et l'on passe aux applications de ces derniers à la résolution d'équations de mailles elliptiques.

§ 2. Résolution de problèmes de différences par la méthode de Fourier

1. Problèmes de différences de valeurs propres pour l'opérateur de Laplace dans un rectangle. Au § 5, ch. I on a étudié les problèmes aux limites de valeurs propres pour l'opérateur de différence divisée seconde, donné sur un tronçon de maillage régulier. Pour deux dimensions, on a pour analogues de ces problèmes des problèmes de valeurs propres associés à l'opérateur de différences de Laplace donné sur un maillage orthogonal régulier dans un rectangle. Profitons de la méthode de séparation des variables pour rechercher les *valeurs propres* λ_k *et les fonctions propres* $\mu_k(i, j)$ *de l'opérateur de différences de Laplace*

$$\Lambda = \Lambda_1 + \Lambda_2, \quad \Lambda_\alpha y = y_{\bar{x}_\alpha x_\alpha}, \quad \alpha = 1, 2.$$

Soit le maillage $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G},\ 0 \leqslant i \leqslant N_1,\ 0 \leqslant j \leqslant N_2,\ h_\alpha N_\alpha = l_\alpha,\ \alpha = 1, 2\}$ donné dans un rectangle $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha,\ \alpha = 1, 2\}$. Désignons, comme habituellement, par ω les nœuds intérieurs et par γ les nœuds de frontière du maillage $\bar{\omega}$.

Le problème de valeurs propres le plus simple pour l'opérateur de Laplace dans le cas des conditions de Dirichlet se pose ainsi: chercher les valeurs du paramètre λ pour lesquelles on a une solution non triviale $y(x)$ du problème suivant:

$$\Lambda y(x) + \lambda y(x) = 0, \quad x \in \omega,$$
$$y(x) = 0, \quad x \in \gamma. \tag{1}$$

Cherchons la fonction propre $\mu_k(i, j)$ du problème (1) correspondant à la valeur λ_k sous la forme

$$\mu_k(i, j) = \mu_{k_1}^{(1)}(i)\,\mu_{k_2}^{(2)}(j), \qquad k = (k_1, k_2). \tag{2}$$

Portons dans (1) au lieu de $y(x_{ij}) = y(i, j)$ la fonction $\mu_k(i, j)$. Vu que

$$\Lambda_1 y(i, j) = \frac{1}{h_1^2}[y(i+1, j) - 2y(i, j) + y(i-1, j)],$$

l'opérateur Λ_1 n'agit que sur la fonction de maille dépendant de l'argument i. De façon analogue, l'opérateur Λ_2 agit sur la fonction dépendant de l'argument j. Aussi après substitution de (2) dans (1) vient-il

$$\mu_{k_2}^{(2)}(j)\,\Lambda_1\mu_{k_1}^{(1)}(i) + \mu_{k_1}^{(1)}(i)\,\Lambda_2\mu_{k_2}^{(2)}(j) + \lambda_k\mu_{k_1}^{(1)}(i)\,\mu_{k_2}^{(2)}(j) = 0 \tag{3}$$

pour $1 \leqslant i \leqslant N_1 - 1$, $1 \leqslant j \leqslant N_2 - 1$, de même que

$$\mu_{k_1}^{(1)}(0) = \mu_{k_1}^{(1)}(N_1) = 0, \qquad \mu_{k_2}^{(2)}(0) = \mu_{k_2}^{(2)}(N_2) = 0. \tag{4}$$

De (3) on déduit que

$$\frac{\Lambda_1\mu_{k_1}^{(1)}(i)}{\mu_{k_1}^{(1)}(i)} = -\frac{\Lambda_2\mu_{k_2}^{(2)}(j)}{\mu_{k_2}^{(2)}(j)} - \lambda_k. \tag{5}$$

Vu que le premier membre ne dépend pas de j, le second membre ne dépend pas également de j. D'autre part, puisque le second membre ne dépend pas de i, le premier membre est également indépendant de i. Donc les deux membres de (5) sont des constantes. Posons

$$\frac{\Lambda_1\mu_{k_1}^{(1)}(i)}{\mu_{k_1}^{(1)}(i)} = -\lambda_{k_1}^{(1)}, \qquad \frac{\Lambda_2\mu_{k_2}^{(2)}(j)}{\mu_{k_2}^{(2)}(j)} = -\lambda_{k_2}^{(2)}, \qquad \lambda_k = \lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)} \tag{6}$$

et ajoutons-y les conditions aux limites (4). On obtient finalement les problèmes de mailles aux valeurs propres unidimensionnels

$$\begin{gathered}\Lambda_1\mu_{k_1}^{(1)} + \lambda_{k_1}^{(1)}\mu_{k_1}^{(1)} = 0, \qquad 1 \leqslant i \leqslant N_1 - 1,\\ \mu_{k_1}^{(1)}(0) = \mu_{k_1}^{(1)}(N_1) = 0\end{gathered} \tag{7}$$

et

$$\begin{gathered}\Lambda_2\mu_{k_2}^{(2)} + \lambda_{k_2}^{(2)}\mu_{k_2}^{(2)} = 0, \qquad 1 \leqslant j \leqslant N_2 - 1,\\ \mu_{k_2}^{(2)}(0) = \mu_{k_2}^{(2)}(N_2) = 0.\end{gathered} \tag{8}$$

Les solutions des problèmes (7) et (8) ont été déjà obtenues au §5 ch. I :

$$\lambda_{k_\alpha}^{(\alpha)} = \frac{4}{h_\alpha^2}\sin^2\frac{k_\alpha\pi}{2N_\alpha} = \frac{4}{h_\alpha^2}\sin^2\frac{k_\alpha\pi h_\alpha}{2l_\alpha}, \qquad k_\alpha = 1, 2, \ldots, N_\alpha - 1,$$

$$\mu_{k_1}^{(1)}(i) = \sqrt{\frac{2}{l_1}}\sin\frac{k_1\pi i}{N_1}, \qquad k_1 = 1, 2, \ldots, N_1 - 1,$$

$$\mu_{k_2}^{(2)}(j) = \sqrt{\frac{2}{l_2}}\sin\frac{k_2\pi j}{N_2}, \qquad k_2 = 1, 2, \ldots, N_2 - 1.$$

Bref, *les fonctions propres et les valeurs propres de l'opérateur de différences de Laplace* Λ sont trouvées pour les conditions aux limites de Dirichlet

$$\mu_k(i,\ j)=\mu_{k_1}^{(1)}(i)\,\mu_{k_2}^{(2)}(j)=\frac{2}{\sqrt{l_1l_2}}\sin\frac{k_1\pi i}{N_1}\sin\frac{k_2\pi j}{N_2},$$

$$0\leqslant i\leqslant N_1,\quad 0\leqslant j\leqslant N_2, \tag{9}$$

$$\lambda_k=\lambda_{k_1}^{(1)}+\lambda_{k_2}^{(2)}=\sum_{\alpha=1}^{2}\frac{4}{h_\alpha^2}\sin^2\frac{k_\alpha\pi h_\alpha}{2l_\alpha},$$

où $k_\alpha=1,\ 2,\ \ldots,\ N_\alpha-1,\ \alpha=1,\ 2.$

Notons les principales propriétés des fonctions propres et des valeurs propres trouvées (9). Introduisons le produit scalaire des fonctions de mailles données sur le maillage $\bar\omega$ de la façon suivante:

$$(u,\ v)=\sum_{x\in\bar\omega}u(x)\,v(x)\,\hbar_1(x_1)\,\hbar_2(x_2),$$

$$\hbar_\alpha(x_\alpha)=\begin{cases}0{,}5h_\alpha, & x_\alpha=0,\ l_\alpha,\\ h_\alpha, & h_\alpha\leqslant x_\alpha\leqslant l_\alpha-h_\alpha.\end{cases}$$

Si l'on pose

$$(u,\ v)_{\bar\omega_\alpha}=\sum_{x_\alpha\in\bar\omega_\alpha}u(x)\,v(x)\,\hbar_\alpha(x_\alpha),\quad \alpha=1,\ 2, \tag{10}$$

où

$$\bar\omega_1=\{x_1(i)=ih_1,\quad 0\leqslant i\leqslant N_1\},\quad \bar\omega_2=\{x_2(j)=jh_2,\ 0\leqslant j\leqslant N_2\},$$

il est vraisemblable que $\bar\omega=\bar\omega_1\times\bar\omega_2$ et $x_{ij}=(x_1(i),\ x_2(j))$, de plus

$$(u,\ v)=((u,\ v)_{\bar\omega_1},\ 1)_{\bar\omega_2}=((u,\ v)_{\bar\omega_2},\ 1)_{\bar\omega_1}. \tag{11}$$

Rappelons qu'au § 5, ch. I on a déjà noté que les fonctions de mailles $\mu_{k_1}^{(1)}(i)$ et $\mu_{k_2}^{(2)}(j)$ sont orthonormées au sens du produit scalaire (10), c.-à-d.

$$(\mu_{k_\alpha}^{(\alpha)},\ \mu_{m_\alpha}^{(\alpha)})_{\bar\omega_\alpha}=\delta_{k_\alpha,\,m_\alpha}=\begin{cases}1, & k_\alpha=m_\alpha,\\ 0, & k_\alpha\neq m_\alpha.\end{cases}$$

Il s'ensuit donc de ce qui précède et de (11) que le système de fonctions propres $\mu_k(i,\ j)$ défini par les formules (9) est orthonormé:

$$(\mu_k,\ \mu_m)=\delta_{k,\,m}=\begin{cases}1, & k=m,\\ 0, & k\neq m,\ k=(k_1,\ k_2),\ m=(m_1,\ m_2).\end{cases}$$

Vu que le nombre de fonctions propres $\mu_k(i,\ j)=\mu_{k_1k_2}(i,\ j)$ vaut $(N_1-1)(N_2-1)$ et coïncide avec celui des nœuds internes

du maillage $\bar{\omega}$, toute fonction de maille $f(i, j)$ donnée sur ω (ou sur $\bar{\omega}$ et devenant nulle sur γ) peut donc être représentée sous la forme suivante:

$$f(i, j) = \sum_{k_1=1}^{N_1-1} \sum_{k_2=1}^{N_2-1} f_{k_1k_2} \mu_{k_1}^{(1)}(i)\, \mu_{k_2}^{(2)}(j), \qquad 1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant j \leqslant N_2 - 1, \tag{12}$$

où les coefficients de Fourier $f_{k_1k_2}$ se définissent ainsi:

$$f_k = f_{k_1k_2} = (f, \mu_k) = \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} f(i, j)\, \mu_{k_1}^{(1)}(i)\, \mu_{k_2}^{(2)}(j)\, h_1 h_2, \qquad k_1 = 1, 2, \ldots, N_1 - 1, \quad k_2 = 1, 2, \ldots, N_2 - 1. \tag{13}$$

Pour les valeurs propres λ_k se vérifie l'estimation

$$\lambda_{\min} = \lambda_1^{(1)} + \lambda_1^{(2)} \leqslant \lambda_k = \lambda_{k_1} + \lambda_{k_2} \leqslant \lambda_{N_1-1}^{(1)} + \lambda_{N_2-1}^{(2)} = \lambda_{\max},$$

où

$$\lambda_{\min} = \sum_{\alpha=1}^{2} \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2l_\alpha} \geqslant 8 \left(\frac{1}{l_1^2} + \frac{1}{l_2^2} \right) > 0,$$

$$\lambda_{\max} = \sum_{\alpha=1}^{2} \frac{4}{h_\alpha^2} \cos^2 \frac{\pi h_\alpha}{2l_\alpha} < 4 \left(\frac{1}{h_1^2} + \frac{1}{h_2^2} \right).$$

Voyons maintenant l'exemple d'un problème de valeurs propres plus compliqué sur l'opérateur de différences de Laplace. Soient toujours données sur les côtés du rectangle les conditions de Dirichlet pour $x_1 = 0$ et $x_1 = l_1$ et les conditions de Neumann pour $x_2 = 0$ et $x_2 = l_2$, c'est-à-dire qu'est posé le problème de valeurs propres suivant:

$$\Lambda y(x) + \lambda y(x) = 0, \quad x \in \omega_1 \times \bar{\omega}_2, \qquad y(x) = 0, \quad x_1 = 0, \quad x_1 = l_1. \tag{14}$$

On a ici $\Lambda = \Lambda_1 + \Lambda_2$, l'opérateur Λ_1 étant défini précédemment, tandis que

$$\Lambda_2 y = \begin{cases} \dfrac{2}{h_2} y_{x_2}, & x_2 = 0, \\ y_{\bar{x}_2 x_2}, & h_2 \leqslant x_2 \leqslant l_2 - h_2, \\ -\dfrac{2}{h_2} y_{\bar{x}_2}, & x_2 = l_2. \end{cases} \tag{15}$$

En utilisant la définition des opérateurs Λ_1 et Λ_2, on peut écrire le problème (14) sous la forme suivante:

$$y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} + \lambda y = 0, \quad x \in \omega,$$

$$\left.\begin{aligned} y_{\bar{x}_1 x_1} + \frac{2}{h_2} y_{x_2} + \lambda y = 0, \quad x_2 = 0, \\ y_{\bar{x}_1 x_1} - \frac{2}{h_2} y_{\bar{x}_2} + \lambda y = 0, \quad x_2 = l_2, \end{aligned}\right\} h_1 \leqslant x_1 \leqslant l_1 - h_1,$$

$$y(0, x_2) = y(l_1, x_2) = 0, \quad 0 \leqslant x_2 \leqslant l_2.$$

On obtient la solution du problème (14) par la méthode de séparation des variables. En portant dans (14) au lieu de y la fonction de maille $\mu_k(i, j)$ tirée de (2), on obtient pour $\mu_{k_1}^{(1)}(i)$ le problème (7) et pour $\mu_{k_2}^{(2)}(j)$ le problème aux limites suivant:

$$\Lambda_2 \mu_{k_2}^{(2)} + \lambda_{k_2}^{(2)} \mu_{k_2}^{(2)} = 0, \quad 0 \leqslant j \leqslant N_2$$

ou, en vertu de (15),

$$\begin{aligned} (\mu_{k_2}^{(2)})_{\bar{x}_2 x_2} + \lambda_{k_2}^{(2)} \mu_{k_2}^{(2)} = 0, \quad & 1 \leqslant j \leqslant N_2 - 1, \\ \frac{2}{h_2} (\mu_{k_2}^{(2)})_{x_2} + \lambda_{k_2}^{(2)} \mu_{k_2}^{(2)} = 0, \quad & j = 0, \\ -\frac{2}{h_2} (\mu_{k_2}^{(2)})_{\bar{x}_2} + \lambda_{k_2}^{(2)} \mu_{k_2}^{(2)} = 0, \quad & j = N_2. \end{aligned} \tag{16}$$

Le problème (16) a aussi été résolu précédemment au § 5, ch. I. La solution est de la forme

$$\lambda_{k_2}^{(2)} = \frac{4}{h_2^2} \sin^2 \frac{k_2 \pi}{2N_2} = \frac{4}{h_2^2} \sin^2 \frac{k_2 \pi h_2}{2l_2}, \quad k_2 = 0, 1, \ldots, N_2,$$

$$\mu_{k_2}^{(2)}(j) = \begin{cases} \sqrt{\frac{2}{l_2}} \cos \frac{k_2 \pi j}{N_2}, & 1 \leqslant k_2 \leqslant N_2 - 1, \\ \sqrt{\frac{1}{l_2}} \cos \frac{k_2 \pi j}{N_2}, & k_2 = 0, N_2. \end{cases} \tag{17}$$

Bref, la solution du problème (14), (15) est trouvée:

$$\mu_k(i, j) = \mu_{k_1}^{(1)}(i) \mu_{k_2}^{(2)}(j), \quad 0 \leqslant i \leqslant N_1, \quad 0 \leqslant j \leqslant N_2,$$

$$\lambda_k = \lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)}, \quad 1 \leqslant k_1 \leqslant N_1 - 1, \quad 0 \leqslant k_2 \leqslant N_2,$$

où $\lambda_{k_1}^{(1)}$ et $\mu_{k_1}^{(1)}(i)$ sont définis plus haut, tandis que $\lambda_{k_2}^{(2)}$ et $\mu_{k_2}^{(2)}(j)$ sont définis dans (17).

De façon analogue sont résolus les problèmes de valeurs propres sur l'opérateur de différences de Laplace dans un rectangle et aux cas d'autres combinaisons de conditions aux limites sur les côtés du rectangle G. La méthode de séparation des variables permet de les réduire aux problèmes unidimensionnels dont les solutions ont

été obtenues au § 5 du chapitre I. La généralisation au cas multidimensionnel est évidente. Rappelons que la solution analytique de problèmes unidimensionnels correspondants sous forme de sinus et de cosinus n'a été obtenue au § 5 du chapitre I que pour les conditions aux limites de première et de seconde espèces, pour leurs combinaisons, ainsi que pour le cas d'un problème aux limites périodique. Aussi si sur les côtés d'un rectangle (sur les faces d'un parallélépipède rectangle dans le cas tridimensionnel) sont données les conditions aux limites mentionnées, les fonctions propres de l'opérateur de différences de Laplace se présentent-elles sous forme de produit des sinus et des cosinus.

2. Equation de Poisson dans un rectangle. Développement en série double. Examinons maintenant la méthode de séparation des variables sous l'angle de son application à la résolution du *problème de différences de Dirichlet pour l'équation de Poisson* donnée sur un maillage régulier dans un rectangle:

$$\begin{gathered}\Lambda y=-\varphi(x), \quad x\in\omega, \quad y(x)=g(x), \quad x\in\gamma,\\ \Lambda=\Lambda_1+\Lambda_2, \quad \Lambda_\alpha y=y_{\bar{x}_\alpha x_\alpha}, \quad \alpha=1,\ 2.\end{gathered} \tag{18}$$

Réduisons d'abord le problème (18) à un problème aux conditions aux limites homogènes en modifiant le second membre de l'équation aux nœuds de la frontière. Le procédé standard de cette transformation consiste à transposer les grandeurs connues dans le second membre de l'équation transcrite au nœud adjacent à la frontière. Par exemple, si $x=(h_1,\ h_2)\in\omega$, l'équation de Poisson est transcrite en ce point sous la forme:

$$\frac{1}{h_1^2}[y(0,\ h_2)-2y(h_1,\ h_2)+y(2h_1,\ h_2)]+\\ +\frac{1}{h_2^2}[y(h_1,\ 0)-2y(h_1,\ h_2)+y(h_1,\ 2h_2)]=-\varphi(h_1,\ h_2).$$

Comme $y(0,\ h_2)=g(0,\ h_2)$, $y(h_1,\ 0)=g(h_1,\ 0)$, en transposant ces grandeurs du premier membre de l'équation dans le second on aura

$$\frac{1}{h_1^2}[-2y(h_1,\ h_2)+y(2h_1,\ h_2)]+\frac{1}{h_2^2}[-2y(h_1,\ h_2)+y(h_1,\ 2h_2)]=\\ =-\left[\varphi(h_1,\ h_2)+\frac{1}{h_1^2}g(0,\ h_2)+\frac{1}{h_2^2}g(h_1,\ 0)\right].$$

En effectuant ces transformations pour chaque point de la frontière, on obtient des équations aux différences qui ne contiennent pas de valeurs de $y(x)$ sur γ dans le premier membre. Les seconds membres des équations des nœuds de la frontière diffèrent du second membre de $\varphi(x)$. Si l'on désigne par $f(x)$ le second membre ainsi

construit, il se détermine au moyen des formules

$$f(x)=\varphi(x)+\frac{1}{h_1^2}\varphi_1(x)+\frac{1}{h_2^2}\varphi_2(x), \qquad x\in\omega, \tag{19}$$

où

$$\varphi_1(x)=\begin{cases} g(0,\ x_2), & x_1=h_1,\\ 0, & 2h_1\leqslant x_1\leqslant l_1-2h_1,\\ g(l_1,\ x_2), & x_1=l_2,\end{cases} \qquad \varphi_2(x)=\begin{cases} g(x_1,\ 0), & x_2=h_2,\\ 0, & 2h_2\leqslant x_2\leqslant l_2-2h_2,\\ g(x_1,\ l_2), & x_2=l_2.\end{cases}$$

Le premier membre des équations transformées diffère pour les nœuds de la frontière de l'écriture de l'opérateur de différences de Laplace. Toutefois, si l'on pose $y(x) = u(x)$, $x \in \omega$, $u(x) = 0$, $x \in \gamma$, les équations en tous les points du maillage ω s'écriront alors de la même façon:

$$\Lambda u = -f(x), \quad x\in\omega, \quad u(x)=0, \quad x\in\gamma. \tag{20}$$

Comme $u(x)$ coïncide avec $y(x)$ pour $x \in \omega$, il suffit de trouver la solution du problème (20).

Cherchons la solution du problème (20). Comme la fonction $u(x)$ devient nulle sur γ, en vertu de ce qui a été dit plus haut, on peut la représenter sous forme d'un développement en fonctions propres $\mu_k(i, j)$ de l'opérateur de Laplace

$$u(i,\ j)=\sum_{k_1=1}^{N_1-1}\sum_{k_2=1}^{N_2-1} u_{k_1k_2}\mu_{k_1}^{(1)}(i)\,\mu_{k_2}^{(2)}(j), \tag{21}$$

ce qui se vérifie pour $0 \leqslant i \leqslant N_1$, $0 \leqslant j \leqslant N_2$. Ensuite, la fonction de maille $f(x)$ donnée sur ω admet également la représentation

$$f(i,\ j)=\sum_{k_1=1}^{N_1-1}\sum_{k_2=1}^{N_2-1} f_{k_1k_2}\mu_{k_1}^{(1)}(i)\,\mu_{k_2}^{(2)}(j) \tag{22}$$

pour $1 \leqslant i \leqslant N_1 - 1$, $1 \leqslant j \leqslant N_2 - 1$, où les coefficients de Fourier $f_{k_1k_2}$ sont définis dans (13). Comme $\mu_k(i, j) = \mu_{k_1}^{(1)}(i)\,\mu_{k_2}^{(2)}(j)$ est la fonction propre de l'opérateur de Laplace correspondant à la valeur propre de λ_k, c'est-à-dire

$$\Lambda\mu_k+\lambda_k\mu_k=0, \qquad x\in\omega, \qquad \lambda_{k_1}^{(1)}+\lambda_{k_2}^{(2)}=\lambda_k,$$

après avoir porté (21) et (22) dans l'équation (20), il vient

$$\Lambda u=\sum_{k_1=1}^{N_1-1}\sum_{k_2=1}^{N_2-1}\left(\lambda_{k_1}^{(1)}+\lambda_{k_2}^{(2)}\right)u_{k_1k_2}\mu_{k_1}^{(1)}(i)\,\mu_{k_2}^{(2)}(j)=-f(i,\ j)=$$

$$=-\sum_{k_1=1}^{N_1-1}\sum_{k_2=1}^{N_2-1} f_{k_1k_2}\mu_{k_1}^{(1)}(i)\,\mu_{k_2}^{(2)}(j),$$

$$1\leqslant i\leqslant N_1-1, \quad 1\leqslant j\leqslant N_2-1.$$

En utilisant les fonctions propres $\mu_k(i, j)$ orthonormées, on obtient de ce qui précède les égalités suivantes:

$$u_{k_1k_2}=\frac{f_{k_1k_2}}{\lambda_{k_1}^{(1)}+\lambda_{k_2}^{(2)}},\qquad 1\leqslant k_1\leqslant N_1-1,\quad 1\leqslant k_2\leqslant N_2-1.$$

En portant cette expression dans (21), on obtient en guise de solution du problème (20) la représentation suivante:

$$u(i,\ j)=\sum_{k_1=1}^{N_1-1}\sum_{k_2=1}^{N_2-1}\frac{f_{k_1k_2}}{\lambda_{k_1}^{(1)}+\lambda_{k_2}^{(2)}}\mu_{k_1}^{(1)}(i)\,\mu_{k_2}^{(2)}(j),\qquad (23)$$
$$0\leqslant i\leqslant N_1,\quad 0\leqslant j\leqslant N_2.$$

Bref, les formules (13) et (23) fournissent la solution du problème (20). Procédons à l'analyse de ces dernières sous l'angle du calcul. Lors du calcul de la solution $u(i, j)$ suivant les formules (13) et (20), où $\mu_k(i, j) = \mu_{k_1}^{(1)}(i)\,\mu_{k_2}^{(2)}(j)$ et $\lambda_k = \lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)}$ sont définis dans (9), il est rationnel d'introduire trois grandeurs auxiliaires: $\varphi_{k_2}(i)$, $\varphi_{k_1k_2}$ et $u_{k_2}(i)$. Dans ce cas le procédé de calcul peut s'organiser ainsi:

$$\varphi_{k_2}(i)=\sum_{j=1}^{N_2-1}f(i,\ j)\sin\frac{k_2\pi j}{N_2},$$
$$1\leqslant k_2\leqslant N_2-1,\quad 1\leqslant i\leqslant N_1-1,\qquad (24)$$

$$\varphi_{k_1k_2}=\sum_{i=1}^{N_1-1}\varphi_{k_2}(i)\sin\frac{k_1\pi i}{N_1},$$
$$1\leqslant k_1\leqslant N_1-1,\quad 1\leqslant k_2\leqslant N_2-1,\qquad (25)$$

$$u_{k_2}(i)=\sum_{k_1=1}^{N_1-1}\frac{\varphi_{k_1k_2}}{\lambda_{k_1}^{(1)}+\lambda_{k_2}^{(2)}}\sin\frac{k_1\pi i}{N_1},$$
$$1\leqslant i\leqslant N_1-1,\quad 1\leqslant k_2\leqslant N_2-1,\qquad (26)$$

$$u(i,\ j)=\frac{4}{N_1N_2}\sum_{k_2=1}^{N_2-1}u_{k_2}(i)\sin\frac{k_2\pi j}{N_2},$$
$$1\leqslant j\leqslant N_2-1,\quad 1\leqslant i\leqslant N_1-1.\qquad (27)$$

Calculons le nombre d'opérations arithmétiques que coûte l'algorithme (24)-(27) en posant que les grandeurs $(\lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)})^{-1}$ sont données et les sommes (24)-(27) se calculent avec l'utilisation de l'algorithme de la transformation rapide de Fourier, décrit au point 2, § 1. Pour l'utilisation de l'algorithme mentionné, il faut admettre que N_1 et N_2 sont des puissances de 2: $N_1 = 2^n$, $N_2 = 2^m$.

Rappelons que les sommes de la forme

$$y_k = \sum_{j=1}^{2^n-1} a_j \sin \frac{k\pi j}{2^n}, \qquad k=1,\ 2,\ \ldots,\ 2^n-1,$$

se calculent avec le coût $Q_+ = (3/2n - 2)\,2^n - n + 2$ additions et soustractions et $Q_* = (n/2 - 1)\,2^n + 1$ multiplications si l'on utilise l'algorithme du point 2, §. 1.

Un calcul élémentaire fournit les coûts suivants en opérations arithmétiques pour le calcul de la solution $u\,(i,\ j)$ suivant les formules (24)-(27):

$$Q_+ = (N_1N_2 - N_1 - N_2)\,[3 \log_2 (N_1N_2) - 8] + \\ + (N_1 + 2) \log_2 N_2 + (N_2 + 2) \log_2 N_1 - 8$$

opérations d'addition et de soustraction et

$$Q_* = (N_1N_2 - N_1 - N_2)\,[\log_2 (N_1N_2) - 2] + N_1 \log_2 N_2 + \\ + N_1 \log_2 N_1 - 2$$

opérations de multiplication. Si l'on néglige les différences entre les opérations arithmétiques, alors, pour $N_1 = N_2 = N = 2^n$, le nombre total d'opérations de l'algorithme (24)-(27) s'élève à

$$Q = (N^2 - 1{,}5N)\,(8 \log_2 N - 10) + 5N + 4 \log_2 N - 10.$$

Ainsi, la méthode décrite de résolution du problème (20) peut être mise en œuvre en $O\,(N^2 \log_2 N)$ opérations arithmétiques. Du même type est l'estimation du nombre d'opérations que coûte la méthode de réduction totale exposée au chapitre III. La confrontation de ces estimations montre que l'algorithme considéré de la méthode de séparation des variables exige 1,5 fois plus d'opérations que la méthode de réduction totale.

Remarquons qu'on peut construire un algorithme analogue au précédent également pour le cas où sur les côtés du rectangle est donnée une combinaison quelconque de conditions aux limites de première ou de seconde espèces et de conditions de périodicité pour lesquelles le problème de différences n'est pas dégénéré. Il est seulement nécessaire de porter dans (13) et (23) les fonctions et les valeurs propres correspondantes, de faire concorder les limites de sommation avec le type des conditions aux limites, ainsi que d'utiliser l'algorithme adéquat de transformation rapide de Fourier du § 1 pour le calcul des sommes ainsi engendrées. L'estimation du nombre d'opérations sera de la même forme que dans le cas du problème de Dirichlet examiné plus haut.

On a décrit la plus simple des variantes de la méthode de séparation des variables. S'il s'agit de résoudre un problème aux limites au sens de différences finies plus général, par exemple, l'équation de Poisson en coordonnées polaires ou cylindriques avec conditions

aux limites admettant la séparation des variables, on peut alors de nouveau utiliser les développements (21) et (22). Mais dans ce cas au moins une des fonctions propres $\mu_{k_1}^{(1)}(i)$ et $\mu_{k_2}^{(2)}(j)$ est différente du sinus ou du cosinus. Cela ne permet pas de recourir à l'algorithme de la transformation rapide de Fourier lors du calcul des sommes nécessaires. Aussi pour ces problèmes le nombre d'opérations arithmétiques sera-t-il du même ordre que dans le cas du calcul direct des sommes qui ne tient pas compte de la forme des fonctions propres $\mu_{k_1}^{(1)}(i)$ et $\mu_{k_2}^{(2)}(j)$, c'est-à-dire est $O(N^3)$.

Il est donc nécessaire de modifier la méthode construite pour que, dans le cas où l'une au moins des fonctions $\mu_{k_1}^{(1)}(i)$ ou $\mu_{k_2}^{(2)}(j)$ est un sinus ou un cosinus, le nombre d'opérations arithmétiques soit une grandeur de l'ordre de $O(N^2 \log_2 N)$. Il va de soi que les problèmes étudiés en ce point peuvent également être résolus au moyen de la méthode modifiée et, comme on le verra plus loin, avec un moindre nombre d'opérations arithmétiques. Cette méthode de développement en série unique sera construite au point 3. Sous l'angle des calculs, elle diffère de celle déjà décrite par le fait que deux des sommes de (24)-(27) peuvent ne pas être calculées et à leur place on résout la série de problèmes aux limites sur les équations aux différences triponctuelles.

3. Développement en série unique. Revenons au problème (20):

$$\Lambda u = -f(x), \quad x \in \omega, \quad u(x) = 0, \quad x \in \gamma,$$
$$\Lambda = \Lambda_1 + \Lambda_2, \quad \Lambda_\alpha u = u_{\bar{x}_\alpha x_\alpha}, \quad \alpha = 1, 2. \tag{28}$$

Etudions la fonction cherchée $u(x_{ij}) = u(i, j)$ et la fonction donnée $f(i, j)$ pour un i fixé, $0 \leqslant i \leqslant N_1$ comme des fonctions de mailles de l'argument j. Comme $u(i, j)$ devient nul pour $j = 0$ et $j = N_2$, et $f(i, j)$ est donnée pour $1 \leqslant j \leqslant N_2 - 1$, on peut les représenter sous forme de sommes de fonctions propres $\mu_{k_2}^{(2)}(j)$ de l'opérateur de différences Λ_2:

$$u(i, j) = \sum_{k_2=1}^{N_2-1} u_{k_2}(i)\,\mu_{k_2}^{(2)}(j), \quad 0 \leqslant j \leqslant N_2, \qquad 0 \leqslant i \leqslant N_1, \tag{29}$$

$$f(i, j) = \sum_{k_2=1}^{N_2-1} f_{k_2}(i)\,\mu_{k_2}^{(2)}(j), \quad 1 \leqslant j \leqslant N_2 - 1, \quad 1 \leqslant i \leqslant N_1 - 1, \tag{30}$$

où

$$\mu_{k_2}^{(2)}(j) = \sqrt{\frac{2}{l_2}} \sin \frac{k_2 \pi j}{N_2}, \quad k_2 = 1, 2, \ldots, N_2 - 1. \tag{31}$$

Portons les expressions (29) et (30) dans (28), compte tenu des égalités

$$\Lambda_2 \mu_{k_2}^{(2)} + \lambda_{k_2}^{(2)} \mu_{k_2}^{(2)} = 0, \quad 1 \leqslant j \leqslant N_2 - 1,$$
$$\mu_{k_2}^{(2)}(0) = \mu_{k_2}^{(2)}(N_2) = 0. \tag{32}$$

Finalement on obtient

$$\sum_{k_2=1}^{N_2-1} [\Lambda_1 u_{k_2}(i) - \lambda_{k_2}^{(2)} u_{k_2}(i) + f_{k_2}(i)]\, \mu_{k_2}^{(2)}(j) = 0$$

pour $1 \leqslant i \leqslant N_1 - 1$, $1 \leqslant j \leqslant N_2 - 1$, de même que $u_{k_2}(0) = u_{k_2}(N_1) = 0$, $k_2 = 1, 2, \ldots, N_2 - 1$.

De là, en raison de l'orthogonalité du système de fonctions propres $\mu_{k_2}^{(2)}(j)$, on obtient une série de problèmes aux limites permettant de déterminer les fonctions $u_{k_2}(i)$, $k_2 = 1, 2, \ldots, N_2 - 1$:

$$\begin{gathered} \Lambda_1 u_{k_2}(i) - \lambda_{k_2}^{(2)} u_{k_2}(i) = -f_{k_2}(i), \quad 1 \leqslant i \leqslant N_1 - 1, \\ u_{k_2}(0) = u_{k_2}(N_1) = 0. \end{gathered} \tag{33}$$

Les valeurs propres $\lambda_{k_2}^{(2)}$ du problème (32) sont connues

$$\lambda_{k_2}^{(2)} = \frac{4}{h_2^2} \sin^2 \frac{k_2 \pi}{2N_2}, \quad k_2 = 1, 2, \ldots, N_2 - 1, \tag{34}$$

tandis que les coefficients de Fourier $f_{k_2}(i)$ pour chaque $1 \leqslant i \leqslant N_1 - 1$ se calculent suivant les formules

$$f_{k_2}(i) = (f, \mu_{k_2}^{(2)})_{\overline{\omega}_2} = \sum_{j=1}^{N_2-1} h_2 f(i, j)\, \mu_{k_2}^{(2)}(j), \quad 1 \leqslant k_2 \leqslant N_2 - 1. \tag{35}$$

Bref, les formules trouvées (29), (31) et (33)-(35) décrivent complètement la méthode de résolution du problème (20). Suivant les formules (35) on obtient pour $1 \leqslant i \leqslant N_1 - 1$ les fonctions $f_{k_2}(i)$, ensuite, pour $1 \leqslant k_2 \leqslant N_2 - 1$, on résout les problèmes (33) pour déterminer les fonctions $u_{k_2}(i)$, tandis qu'à l'aide des formules (29) on calcule la solution $u(i, j)$ cherchée.

Examinons maintenant l'algorithme mettant en œuvre la méthode décrite. Au lieu de $u_{k_2}(i)$ et $f_{k_2}(i)$ il est commode d'introduire de nouvelles fonctions auxiliaires $v_{k_2}(i)$ et $\varphi_{k_2}(i)$ suivant les formules

$$u_{k_2}(i) = \frac{\sqrt{2l_2}}{N_2} v_{k_2}(i), \quad f_{k_2}(i) = \frac{\sqrt{2l_2}}{N_2} \varphi_{k_2}(i). \tag{36}$$

Portons (31) et (36) dans (29), (33) et (35), tenons compte de ce que $h_2 N_2 = l_2$ et répartissons l'opérateur de différences Λ_1 entre les points. Finalement on obtient

$$\varphi_{k_2}(i) = \sum_{j=1}^{N_2-1} f(i, j) \sin \frac{k_2 \pi j}{N_2}, \quad \left. \begin{array}{l} 1 \leqslant k_2 \leqslant N_2 - 1, \\ 1 \leqslant i \leqslant N_1 - 1, \end{array} \right\} \tag{37}$$

$$\left. \begin{array}{l} -v_{k_2}(i-1) + (2 + h_1^2 \lambda_{k_2}^{(2)})\, v_{k_2}(i) - v_{k_2}(i+1) = h_1^2 \varphi_{k_2}(i), \\ 1 \leqslant i \leqslant N_1 - 1, \quad v_{k_2}(0) = v_{k_2}(N_1) = 0, \quad 1 \leqslant k_2 \leqslant N_2 - 1, \end{array} \right\} \tag{38}$$

$$u(i, j) = \frac{2}{N_2} \sum_{k_2=1}^{N_2-1} v_{k_2}(i) \sin \frac{k_2 \pi j}{N_2}, \quad \left. \begin{array}{l} 1 \leqslant j \leqslant N_2 - 1, \\ 1 \leqslant i \leqslant N_1 - 1, \end{array} \right\} \tag{39}$$

où $\lambda_{k_2}^{(2)}$ est défini dans (34).

Les sommes (37) et (39) doivent apparemment être calculées en utilisant l'algorithme de la transformation rapide de Fourier exposé au point 2, § 1. Pour résoudre les problèmes aux limites triponctuels (38) il est logique d'utiliser l'algorithme de balayage construit au § 1, chapitre II. Pour le problème (38) l'algorithme de balayage est décrit par les formules

$$\begin{aligned} \alpha_{i+1} &= \frac{1}{c_{k_2}-\alpha_i}, & 1\leqslant i\leqslant N_1-1, \quad & \alpha_1=0, \\ \beta_{i+1} &= [h_1^2\varphi_{k_2}(i)+\beta_i]\,\alpha_{i+1}, & 1\leqslant i\leqslant N_1-1, \quad & \beta_1=0, \\ v_{k_2}(i) &= \alpha_{i+1}v_{k_2}(i+1)+\beta_{i+1}, & 1\leqslant i\leqslant N_1-1, \quad & v_{k_2}(N_1)=0, \end{aligned} \tag{40}$$

où $c_{k_2}=2+h_1^2\lambda_{k_2}^{(2)}$ et $k_2=1,\ 2,\ \ldots,\ N_2-1$.

Comparons les formules (37), (39) et (40) avec les formules (24)-(27) obtenues auparavant pour la méthode de développement en série double. Au lieu de calculer les deux sommes (25) et (26), on résout la série des problèmes aux limites (38) par la méthode du balayage (40). Aussi le calcul des sommes (37) et (39) coûtera-t-il à peu près la moitié des opérations arithmétiques de l'algorithme (24)-(27). Le coût complémentaire, dû à la résolution du problème (38), montera apparemment à $O(N_1N_2)$ opérations, mais sera sans effet sur le terme principal de l'estimation du nombre d'opérations arithmétiques de l'algorithme (37), (39), (40). Donnons les estimations précises du nombre d'opérations de cet algorithme. On a (pour $N_2=2^m$) $Q_\pm = [(3\log_2 N_2-1)\,N_2-2\log_2 N_2+1]\,(N_1-1)$ additions et soustractions, $Q_* = [(\log_2 N_2+2)\,N_2-2]\,(N_1-1)$ multiplications et $Q_{,} = (N_1-1)\,(N_2-1)$ divisions et, pour $N_1=N_2=N=2^n$, le nombre d'opérations s'élève au total à

$$Q = (N^2-1{,}5N)\,(4\log_2 N+2)-N+2\log_2 N+2.$$

On a examiné la méthode de développement en série unique sur la base de l'exemple du problème discret de Dirichlet pour l'équation de Poisson. Le point essentiel est le fait que les fonctions propres de l'opérateur de différences Λ_2 admettent l'utilisation de l'algorithme de la transformation rapide de Fourier pour le calcul des sommes correspondantes. Cette éventualité se présentera également pour le cas où sur les côtés $x_2=0$ et $x_2=l_2$ du rectangle $\overline{G}$ sont données, au lieu des conditions aux limites de première espèce, les conditions de seconde espèce ou la combinaison des conditions de première et de seconde espèces, de même que pour le cas de conditions périodiques.

Voyons en guise d'exemple le problème aux limites pour l'équation de Poisson suivant:

$$u_{\bar{x}_1x_1}+u_{\bar{x}_2x_2}=-\varphi(x), \quad x\in\omega,$$

$$u(x)=0, \quad x_1=0,\ l_1, \quad 0\leqslant x_2\leqslant l_2,$$

$$u_{\bar{x}_1x_1}+\frac{2}{h_2}u_{x_2}=-\varphi(x)-\frac{2}{h_2}g_{-2}(x), \quad x_2=0, \tag{41}$$

$$u_{\bar{x}_1x_1}-\frac{2}{h_2}u_{\bar{x}_2}=-\varphi(x)-\frac{2}{h_2}g_{+2}(x), \quad x_2=l_2,$$

$$h_1\leqslant x_1\leqslant l_2-h_1.$$

Le schéma (41) est l'approximation au sens de différences finies du problème

$$\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}=-\varphi(x), \quad x\in G,$$

$$u(x)=0, \quad x_1=0,\ l_1, \quad 0\leqslant x_2\leqslant l_2,$$

$$\frac{\partial u}{\partial x_2}=-g_{-2}(x), \quad x_2=0,$$

$$-\frac{\partial u}{\partial x_2}=-g_{+2}(x), \quad x_2=l_2, \quad 0\leqslant x_1\leqslant l_1.$$

Ecrivons le problème (41) sous une autre forme en posant

$$\Lambda_2 u=\begin{cases}\dfrac{2}{h_2}u_{x_2}, & x_2=0,\\ u_{\bar{x}_2x_2}, & h_2\leqslant x_2\leqslant l_2-h_2,\\ -\dfrac{2}{h_2}u_{\bar{x}_2}, & x_2=l_2,\end{cases}$$

$$\varphi_2(x)=\begin{cases}\dfrac{2}{h_2}g_{-2}(x), & x_2=0,\\ 0, & h_2\leqslant x_2\leqslant l_2-h_2,\\ \dfrac{2}{h_2}g_{+2}(x), & x_2=l_2,\end{cases}$$

$$f(x)=\varphi(x)+\varphi_2(x), \quad \Lambda_1 u=u_{\bar{x}_1x_1}$$

pour $h_1\leqslant x_1\leqslant l_1-h_1$, $0\leqslant x_2\leqslant l_2$.

En nouvelles notations le problème (41) s'écrira sous la forme

$$\Lambda u=(\Lambda_1+\Lambda_2)\,u=-f(x), \quad h_1\leqslant x_1\leqslant l_1-h_1, \quad 0\leqslant x_2\leqslant l_2,$$
$$u(x)=0, \quad x_1=0,\ l_1, \quad 0\leqslant x_2\leqslant l_2. \tag{42}$$

En décomposant $u(i, j)$ et $f(i, j)$ en sommes de fonctions propres de l'opérateur Λ_2, il vient

$$u(i, j)=\sum_{k_2=0}^{N_2}u_{k_2}(i)\,\mu_{k_2}^{(2)}(j), \quad 0\leqslant j\leqslant N_2, \quad 0\leqslant i\leqslant N_1,$$
$$f(i, j)=\sum_{k_2=0}^{N_2}f_{k_2}(i)\,\mu_{k_2}^{(2)}(j), \quad 0\leqslant j\leqslant N_2, \quad 1\leqslant i\leqslant N_1-1, \tag{43}$$

où

$$\mu_{k_2}^{(2)}(j)=\begin{cases}\sqrt{\frac{1}{l_2}}\cos\frac{k_2\pi j}{N_2}, & k_2=0,\ N_2,\\ \sqrt{\frac{2}{l_2}}\cos\frac{k_2\pi j}{N_2}, & 1\leqslant k_2\leqslant N_2-1\end{cases}$$

est la fonction propre de l'opérateur Λ_2 correspondant à la valeur propre

$$\lambda_{k_2}^{(2)}=\frac{4}{h_2^2}\sin^2\frac{k_2\pi}{2N_2},\quad k_2=0,\ 1,\ \ldots,\ N_2. \tag{44}$$

Le coefficient de Fourier $f_{k_2}(i)$ pour chaque $1\leqslant i\leqslant N_1-1$ se calcule suivant les formules

$$f_{k_2}(i)=\sum_{j=1}^{N_2-1}h_2 f(i,\ j)\,\mu_{k_2}^{(2)}(j)+0{,}5h_2\,[f(i,\ 0)\,\mu_{k_2}^{(2)}(0)+f(i,\ N_2)\,\mu_{k_2}^{(2)}(N_2)].$$

En portant (43) dans (42) on obtient pour le problème considéré (42) l'analogue suivant des formules (37)-(39):

$$\varphi_{k_2}(i)=\sum_{j=0}^{N_2}\rho_j f(i,\ j)\cos\frac{k_2\pi j}{N_2},$$

$$0\leqslant k_2\leqslant N_2,\quad 1\leqslant i\leqslant N_1-1,$$

$$-v_{k_2}(i-1)+(2+h_1^2\lambda_{k_2}^{(2)})\,v_{k_2}(i)-v_{k_2}(i+1)=h_1^2\varphi_{k_2}(i),$$

$$1\leqslant i\leqslant N_1-1,\quad v_{k_2}(0)=v_{k_2}(N_1)=0,\quad 0\leqslant k_2\leqslant N_2,$$

$$u(i,\ j)=\frac{2}{N_2}\sum_{k_2=0}^{N_2}\rho_{k_2}v_{k_2}(i)\cos\frac{k_2\pi j}{N_2},$$

$$0\leqslant j\leqslant N_2,\quad 1\leqslant i\leqslant N_1-1,$$

où $\lambda_{k_2}^{(2)}$ est défini dans (44), et

$$\rho_j=\begin{cases}0{,}5, & j=0,\ N_2,\\ 1, & 1\leqslant j\leqslant N_2-1.\end{cases}$$

Procédons à l'estimation du nombre d'opérations qu'implique la construction de l'algorithme pour $N_1=N_2=N=2^n$: $Q_\pm=$ $=[(3\log_2 N_2-1)\,N_2+2\log_2 N_2+7]\,(N_1-1)$ additions et soustractions, $Q_*=[(\log_2 N_2+2)\,N_2+10]\,(N_1-1)$ multiplications et $Q_/=(N_2+1)\,(N_1-1)$ divisions, en tout

$$Q=\left(N^2-\frac{N}{2}\right)(4\log_2 N+2)+17N-2\log_2 N-18.$$

Ensuite, vu que dans la méthode de développement en série unique des fonctions propres de l'opérateur de différences Λ_1 ne sont pas utilisées et la seule exigence envers Λ_1 est la possibilité de sépa-

ration des variables, on peut en qualité de Λ_1 choisir un opérateur plus général que celui qu'on a considéré. Si l'on se limite aux équations elliptiques du second ordre, au cas le plus général de choix de l'opérateur Λ_1 correspond l'approximation de l'opérateur différentiel au sens de différences finies

$$L_1u=\frac{1}{k_2(x_1)}\frac{\partial}{\partial x_1}\left(k_1(x_1)\frac{\partial u}{\partial x_1}\right)+r(x_1)\frac{\partial u}{\partial x_1}-q(x_1)u,$$

dont les coefficients ne dépendent que de x_1. Quant aux conditions aux limites sur les côtés $x_1 = 0$ et $x_1 = l_2$ du rectangle $\bar{G}$, elles peuvent être une combinaison quelconque des conditions aux limites de première, de seconde ou de troisième espèce (les coefficients de la condition aux limites de troisième espèce doivent être des constantes). Cela permet de résoudre les problèmes aux limites pour l'équation de Poisson en coordonnées cylindriques, sphériques et polaires.

§ 3. Méthode de réduction non totale

1. Combinaison des méthodes de Fourier et de réduction. La méthode de développement en série unique construite au point 3, § 2 a permis de se limiter au calcul de deux sommes de Fourier en $O(N_1N_2\log_2 N_2)$ opérations et à la résolution d'une série de problèmes aux limites triponctuels en $O(N_1N_2)$ opérations. Apparemment, la perfection subséquente de la méthode de séparation des variables est possible dans la voie de diminution de termes des sommes calculées avec éventualité d'utilisation de l'algorithme de transformation rapide de Fourier.

Ce but pourra être atteint par combinaison de la méthode de développement en série unique avec la méthode de réduction étudiée au chapitre III. Construisons d'abord une méthode combinée pour le plus simple des problèmes de Dirichlet

$$\begin{gathered}\Lambda u=-f(x),\quad x\in\omega,\quad u(x)=0,\quad x\in\gamma,\\ \Lambda=\Lambda_1+\Lambda_2,\quad \Lambda_\alpha u=u_{\bar{x}_\alpha x_\alpha},\quad \alpha=1,2\end{gathered}\tag{1}$$

sur un maillage rectangulaire $\bar{\omega}$.

Pour simplifier la description de la méthode, passons de l'écriture ponctuelle (scalaire) du problème (1) à l'écriture vectorielle.

Introduisons le vecteur des inconnues U_j de la façon suivante:

$$U_j=(u(1,j),\ u(2,j),\ \ldots,\ u(N_1-1,j)),\quad 0\leqslant j\leqslant N_2,$$

et définissons le vecteur des seconds membres F_j à l'aide de la formule

$$F_j=(h_2^2f(1,j),\ h_2^2f(2,j),\ \ldots,\ h_2^2f(N_1-1,j)),\quad 1\leqslant j\leqslant N_2-1.$$

Le problème de différences (1) peut être alors écrit (voir ch. III, § 1) sous l'aspect du système suivant d'équations vectorielles:

$$\begin{aligned} &-U_{j-1} + CU_j - U_{j+1} = F_j, \quad 1 \leqslant j \leqslant N_2 - 1, \\ &U_0 = U_{N_2} = 0, \end{aligned} \tag{2}$$

où la matrice carrée tridiagonale C est définie par les égalités

$$CU_j = ((2E - h_2^2\Lambda_1)\, u\,(1,\, j),\ \ldots,\ (2E - h_2^2\Lambda_1)\, u\,(N_1 - 1,\, j)),$$
$$\Lambda_1 u = u_{\bar{x}_1 x_1}, \quad u\,(0,\, j) = u\,(N_1,\, j) = 0.$$

Soit N_2 la puissance de $2: N_2 = 2^m$. Rappelons que dans la méthode de réduction totale (voir ch. III, § 2) le premier pas du procédé d'élimination consiste à dégager de (2) le système « raccourci » d'inconnues U_j aux numéros j pairs

$$\begin{aligned} &-U_{j-2} + C^{(1)}U_j - U_{j+2} = F_j^{(1)}, \quad j = 2,\, 4,\, 6,\, \ldots,\, N_2 - 2, \\ &U_0 = U_{N_2} = 0 \end{aligned} \tag{3}$$

et d'équations

$$CU_j = F_j + U_{j-1} + U_{j+1}, \quad j = 1,\, 3,\, 5,\, \ldots,\, N_2 - 1 \tag{4}$$

permettant de déterminer les inconnues aux numéros j impairs. On a posé ici

$$F_j^{(1)} = F_{j-1} + CF_j + F_{j+1}, \quad j = 2,\, 4,\, 6,\, \ldots,\, N_2 - 2, \tag{5}$$

$$C^{(1)} = [C]^2 - 2E. \tag{6}$$

Occupons-nous du système (3). Posons

$$V_j = (v\,(1,\, j),\ v\,(2,\, j),\ \ldots,\ v\,(N_1 - 1,\, j)),$$

$$\Phi_j = (h_2^2\varphi\,(1,\, j),\ h_2^2\varphi\,(2,\, j),\ \ldots,\ h_2^2\varphi\,(N_1 - 1,\, j))$$

et supposons que

$$V_j = U_{2j}, \quad 0 \leqslant j \leqslant N_2/2, \quad \Phi_j = F_{2j}^{(1)}, \quad 1 \leqslant j \leqslant N_2/2 - 1.$$
$$v\,(0,\, j) = v\,(N_1,\, j) = 0, \quad 0 \leqslant j \leqslant N_2/2.$$

Ces notations permettent d'écrire le système (3) sous la forme

$$\begin{aligned} &-V_{j-1} + C^{(1)}V_j - V_{j+1} = \Phi_j, \quad j = 1,\, 2,\, \ldots,\, M_2 - 1, \\ &V_0 = V_{M_2} = 0, \end{aligned} \tag{7}$$

où $2M_2 = N_2$ et, en vertu de (5),

$$\Phi_j = F_{2j-1} + CF_{2j} + F_{2j+1}, \quad j = 1,\, 2,\, \ldots,\, M_2 - 1. \tag{8}$$

Remarquons maintenant que la fonction de maille $v\,(i,\, j)$ est définie pour $0 \leqslant i \leqslant N_1$ et $0 \leqslant j \leqslant M_2$ et devient nulle pour $j = 0$ et $j = M_2$. La fonction $\varphi\,(i,\, j)$ est définie pour $1 \leqslant i \leqslant$ $\leqslant N_1 - 1$ et $1 \leqslant j \leqslant M_2 - 1$. Aussi ces fonctions peuvent-elles

se représenter sous forme de séries uniques de Fourier

$$\begin{aligned} v(i, j) &= \sum_{k_2=1}^{M_2-1} y_{k_2}(i)\,\mu_{k_2}^{(2)}(j), \quad 0 \leqslant i \leqslant N_1, \quad 0 \leqslant j \leqslant M_2, \\ \varphi(i, j) &= \sum_{k_2=1}^{M_2-1} z_{k_2}(i)\,\mu_{k_2}^{(2)}(j), \\ & \qquad 1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant j \leqslant M_2 - 1, \end{aligned} \tag{9}$$

où les fonctions

$$\mu_{k_2}^{(2)}(j) = \frac{2}{\sqrt{l_2}} \sin \frac{k_2 \pi j}{M_2}, \quad k_2 = 1, 2, \ldots, M_2 - 1 \tag{10}$$

forment un système orthonormé sur le maillage $\bar{\omega}$ au sens du produit scalaire

$$(u, v) = \sum_{j=1}^{M_2-1} u(j)\,v(j)\,h_2.$$

Les coefficients de Fourier $z_{k_2}(i)$ de la fonction $\varphi(i, j)$ s'obtiennent suivant les formules

$$\begin{aligned} z_{k_2}(i) = (\varphi, \mu_{k_2}^{(2)}) &= \sum_{j=1}^{M_2-1} h_2 \varphi(i, j)\,\mu_{k_2}^{(2)}(j), \\ 1 \leqslant k_2 \leqslant M_2 - 1, &\quad 1 \leqslant i \leqslant N_1 - 1. \end{aligned} \tag{11}$$

A partir de (9) on obtient pour les vecteurs V_j et Φ_j les développements suivants:

$$\begin{aligned} V_j &= \sum_{k_2=1}^{M_2-1} Y_{k_2} \mu_{k_2}^{(2)}(j), \qquad 0 \leqslant j \leqslant M_2, \\ \Phi_j &= \sum_{k_2=1}^{M_2-1} h_2^2 Z_{k_2} \mu_{k_2}^{(2)}(j), \quad 1 \leqslant j \leqslant M_2 - 1, \end{aligned} \tag{12}$$

où

$$\begin{aligned} Y_{k_2} &= (y_{k_2}(1), y_{k_2}(2), \ldots, y_{k_2}(N_1 - 1)), \\ Z_{k_2} &= (z_{k_2}(1), z_{k_2}(2), \ldots, z_{k_2}(N_1 - 1)). \end{aligned}$$

Portons (12) dans (7) et tenons compte de l'égalité

$$\mu_{k_2}^{(2)}(j-1) + \mu_{k_2}^{(2)}(j+1) = 2 \cos \frac{k_2 \pi}{M_2} \mu_{k_2}^{(2)}(j), \quad 1 \leqslant k_2 \leqslant M_2 - 1.$$

Il vient

$$\sum_{k_2=1}^{M_2-1} \left(C^{(1)} - 2 \cos \frac{k_2 \pi}{M_2} E \right) Y_{k_2} \mu_{k_2}^{(2)}(j) = \sum_{k_2=1}^{M_2-1} h_2^2 Z_{k_2} \mu_{k_2}^{(2)}(j),$$

d'où, en vertu de l'orthonormalité du système (10), on aura

$$\left(C^{(1)}-2\cos\frac{k_2\pi}{M_2}E\right)Y_{k_2}=h_2^2Z_{k_2},\quad 1\leqslant k_2\leqslant M_2-1. \qquad (13)$$

Utilisons la relation (6) et il vient

$$C^{(1)}-2\cos\frac{k_2\pi}{M_2}E=[C]^2-2\left(1+\cos\frac{k_2\pi}{M_2}\right)E=$$
$$=\left(C-2\cos\frac{k_2\pi}{2M_2}E\right)\left(C+2\cos\frac{k_2\pi}{2M_2}E\right).$$

Comme la matrice $C^{(1)}-2\cos\frac{k_2\pi}{M_2}E$ est factorisée, pour résoudre l'équation (13) on peut utiliser l'algorithme

$$\begin{gathered}\left(C-2\cos\frac{k_2\pi}{2M_2}E\right)W_{k_2}=h_2^2Z_{k_2},\\ \left(C+2\cos\frac{k_2\pi}{2M_2}E\right)Y_{k_2}=W_{k_2},\quad 1\leqslant k_2\leqslant M_2-1,\end{gathered} \qquad (14)$$

où le vecteur auxiliaire W_{k_2} comprend les composantes $w_{k_2}(i)$:

$$W_{k_2}=(w_{k_2}(1),\ w_{k_2}(2),\ \ldots,\ w_{k_2}(N_1-1)),$$
$$w_{k_2}(0)=w_{k_2}(N_1)=0.$$

Les formules cherchées sont ainsi obtenues. En passant dans (4), (8) et (14) de l'écriture vectorielle à l'écriture scalaire et utilisant la relation $u(i, 2j)=v(i, j)$, tirée de la définition de V_j, on obtient les formules suivantes permettant de construire la méthode:

$$\begin{gathered}\varphi(i, j)=f(i, 2j-1)+2f(i, 2j)+f(i, 2j+1)-h_2^2\Lambda_1 f(i, 2j),\\ 1\leqslant j\leqslant N_2/2-1,\quad 1\leqslant i\leqslant N_1-1,\quad f(0, 2j)=f(N_1, 2j)=0\end{gathered} \qquad (15)$$

pour le calcul de la fonction $\varphi(i, j)$; les équations

$$\begin{gathered}2\left(1-\cos\frac{k_2\pi}{2M_2}\right)w_{k_2}(i)-h_2^2\Lambda_1 w_{k_2}(i)=h_2^2 z_{k_2}(i),\\ 1\leqslant i\leqslant N_1-1,\\ w_{k_2}(0)=w_{k_2}(N_1)=0,\\ 2\left(1+\cos\frac{k_2\pi}{2M_2}\right)y_{k_2}(i)-h_2^2\Lambda_1 y_{k_2}(i)=w_{k_2}(i),\\ 1\leqslant i\leqslant N_1-1,\\ y_{k_2}(0)=y_{k_2}(N_1)=0\end{gathered} \qquad (16)$$

pour la détermination de $y_{k_2}(i)$ pour $k_2=1, 2, \ldots, M_2-1$ et les équations

$$\begin{gathered}2u(i, 2j-1)-h_2^2\Lambda_1 u(i, 2j-1)=\\ =h_2^2 f(i, 2j-1)+u(i, 2j-2)+u(i, 2j),\\ 1\leqslant i\leqslant N_1-1,\quad u(0, 2j-1)=u(N_1, 2j-1)=0\end{gathered} \qquad (17)$$

permettant de trouver la solution pour $j = 1, 2, \ldots, M_2$. Pour les coefficients de Fourier $z_{k_2}(i)$ on a la formule (11), et à partir de (9) on obtient

$$u(i, 2j) = \sum_{k_2=1}^{M_2-1} y_{k_2}(i)\,\mu_{k_2}^{(2)}(j), \quad 1 \leqslant j \leqslant M_2 - 1, \quad 1 \leqslant i \leqslant N_1 - 1. \tag{18}$$

Bref, les formules (10), (11), (15)-(18) décrivent complètement la méthode de résolution du problème (1) qui constitue une combinaison de la méthode de développement en série unique de Fourier et de la méthode de réduction.

Passons maintenant à la construction de l'algorithme de la méthode. Dans les formules (9), (16) et (18) procédons à la substitution $y_{k_2}(i) = a\bar{y}_{k_2}(i)$, $w_{k_2}(i) = a\bar{w}_{k_2}(i)$, $z_{k_2}(i) = a\bar{z}_{k_2}(i)$, où $a = 2\sqrt{l_2}/N_2$ et dans les formules ainsi obtenues laissons tomber la barre. Cette substitution permet de se passer du facteur de normalisation $2/\sqrt{l_2}$ accompagnant la fonction propre $\mu_{k_2}^{(2)}(j)$ dans les sommes (11) et (18). Ensuite, les problèmes (16) et (17) seront résolus par la méthode du balayage. On se convainc sans peine que les conditions de correction et de stabilité de la méthode du balayage ordinaire sont ici remplies. Notons la singularité du problème (17). Vu que les coefficients de l'équation (17) sont indépendants de j, il est nécessaire de calculer les coefficients de balayage α_i une seule fois avec la résolution du problème (17) pour $j = 1$ et d'utiliser ensuite ces derniers pour la résolution des équations (17) pour des j restants.

Donnons les formules de calculs utilisées. On calcule d'abord

$$\varphi(i, j) = f(i, 2j-1) + f(i, 2j+1) + 2\left(1 + \frac{h_2^2}{h_1^2}\right) f(i, 2j) - \\ - \frac{h_2^2}{h_1^2}\left[f(i-1, 2j) + f(i+1, 2j)\right], \tag{19}$$

$$1 \leqslant j \leqslant M_2 - 1, \quad 1 \leqslant i \leqslant N_1 - 1,$$

où $f(0, 2j) = f(N_1, 2j) = 0$. Les valeurs $\varphi(i, j)$ peuvent prendre la place de $f(i, 2j)$. Les sommes

$$z_{k_2}(i) = \sum_{j=1}^{M_2-1} \varphi(i, j) \sin\frac{k_2\pi j}{M_2}, \quad 1 \leqslant k_2 \leqslant M_2 - 1 \tag{20}$$

pour $1 \leqslant i \leqslant N_1 - 1$ se calculent à l'aide de l'algorithme de la transformation rapide discrète de Fourier, et $z_{k_2}(i)$ prend la place

de $\varphi(i, k_2)$. Au moyen de la méthode du balayage

$$\alpha_{i+1} = 1/(c_{k_2} - \alpha_i), \quad \beta_{i+1} = [h_1^2 z_{k_2}(i) + \beta_i]\, \alpha_{i+1},$$
$$i = 1, 2, \ldots, N-1, \quad \alpha_1 = \beta_1 = 0,$$
$$w_{k_2}(i) = \alpha_{i+1} w_{k_2}(i+1) + \beta_{i+1}, \quad i = N_1 - 1, N_1 - 2, \ldots, 1, \tag{21}$$
$$w_{k_2}(N_1) = 0, \quad c_{k_2} = 2 + 2\frac{h_1^2}{h_2^2} - 2\frac{h_1^2}{h_2^2}\cos\frac{k_2\pi}{N_2}$$

se résout la première des équations (16) et, de façon analogue, suivant les formules

$$\alpha_{i+1} = \frac{1}{c_{k_2} - \alpha_i}, \quad \beta_{i+1} = \left[\frac{h_1^2}{h_2^2} w_{k_2}(i) + \beta_i\right] \alpha_{i+1},$$
$$i = 1, 2, \ldots, N_1 - 1, \quad \alpha_1 = \beta_1 = 0,$$
$$y_{k_2}(i) = \alpha_{i+1} y_{k_2}(i+1) + \beta_{i+1}, \quad i = N_1 - 1, N_2 - 1, \ldots, 1, \tag{22}$$
$$y_{k_2}(N_1) = 0, \quad c_{k_2} = 2 + 2\frac{h_1^2}{h_2^2} + 2\frac{h_1^2}{h_2^2}\cos\frac{k_2\pi}{N_2},$$

est résolue la seconde des équations (16). Le calcul s'effectue ici de proche en proche pour $k_2 = 1, 2, \ldots, M_2 - 1$ et les résultats $w_{k_2}(i)$ et $y_{k_2}(i)$ prennent successivement la place de $z_{k_2}(i)$.

Pour le calcul des sommes

$$u(i, 2j) = \frac{4}{N_2} \sum_{k_2=1}^{M_2-1} y_{k_2}(i) \sin\frac{k_2\pi j}{M_2}, \quad 1 \leqslant j \leqslant M_2 - 1, \tag{23}$$

pour $1 \leqslant i \leqslant N_1 - 1$ on utilise de nouveau l'algorithme de la transformation rapide de Fourier. Les problèmes (17) sont résolus au moyen de la méthode du balayage compte tenu de la singularité notée de ces équations:

$$\alpha_{i+1} = 1/(c - \alpha_i), \quad i = 1, 2, \ldots, N_1 - 1, \quad \alpha_1 = 0,$$
$$\beta_{i+1} = \left[h_1^2 f(i, 2j-1) + \frac{h_1^2}{h_2^2}(u(i, 2j-2) + u(i, 2j)) + \beta_i\right] \alpha_{i+1},$$
$$i = 1, 2, \ldots, N_1 - 1, \quad \beta_1 = 0, \tag{24}$$
$$u(i, 2j-1) = \alpha_{i+1} u(i+1, 2j-1) + \beta_{i+1},$$
$$i = N_1 - 1, N_1 - 2, \ldots, 1, \quad u(N_1, 2j-1) = 0,$$
$$c = 2(1 + h_1^2/h_2^2)$$

pour $1 \leqslant j \leqslant M_2$. La solution $u(i, j)$ se dispose à la place de $f(i, j)$ et, par suite, l'algorithme peut se passer de mémoire complémentaire pour l'information intermédiaire.

Calculons le nombre d'opérations arithmétiques de la mise en œuvre de l'algorithme (19)-(24). Le calcul suivant les formules (19), (21), (22) et (24) coûte $Q_\pm = (6{,}5N_2 - 9)(N_1 - 1)$ additions et

soustractions, $Q_* = (6N_2 - 8)(N_1 - 1)$ multiplications et $Q_, = (N_2 - 1)(N_1 - 1)$ divisions. Pour le calcul des sommes (20) et (23) il faut

$$Q_\pm = \left[\left(\frac{3}{2}\log_2 N_2 - \frac{7}{2}\right) N_2 - 2\log_2 N_2 + 6\right](N_1 - 1)$$

additions et soustractions et

$$Q_* = \left[\left(\frac{1}{2}\log_2 N_2 - 1\right) N_2 + 1\right](N_1 - 1)$$

multiplications. En tout pour $N_1 = N_2 = N = 2^n$ l'algorithme (19)-(24) coûte

$$Q = (N^2 - 2N)(2\log_2 N + 9) - 2N + 2\log_2 N + 11 \quad (25)$$

opérations arithmétiques.

A titre de comparaison, donnons le nombre d'opérations de la méthode de développement en série unique (voir point 3, § 2)

$$Q = \left(N^2 - \frac{3}{2}N\right)(4\log_2 N + 2) - N + 2\log_2 N + 2, \quad (26)$$

le nombre de la méthode de développement en série double (voir point 2, § 2)

$$Q = \left(N^2 - \frac{3}{2}N\right)(8\log_2 N - 10) + 5N + 4\log_2 N - 10, \quad (27)$$

ainsi que le nombre d'opérations du second algorithme de la méthode de réduction totale (voir ch. III, § 2, point 4):

$$Q = \left(N^2 - \frac{11}{5}N\right)(5\log_2 N + 5) + N + 6\log_2 N + 5. \quad (28)$$

Si l'on compare dans les estimations (25)-(28) les constantes associées au terme principal $N^2 \log_2 N$, il apparaît que la méthode combinée exige à peu près 4 fois moins d'opérations arithmétiques que la méthode de développement en série double. Cette conclusion se vérifie pour des N grands. Afin d'obtenir des relations réelles entre les méthodes confrontées pour des N admissibles, donnons le tableau des valeurs de Q pour ces méthodes.

Tableau 4

Estimation / N	(25)	(26)	(27)	(28)
32	18 383	21 496	29 510	28 541
64	83 601	104 950	152 334	138 537
128	371 515	485 708	745 582	643 921

En résumé, la combinaison des méthodes de Fourier et de réduction permet de réduire le nombre d'opérations devant la méthode de départ de développement en série unique. Généralisons cette

méthode combinée en y incluant l opérations d'élimination de la méthode de réduction avant d'aborder le développement en série unique. On peut alors considérer la méthode du point 3, § 2 comme un cas particulier de la méthode généralisée avec $l = 0$, la méthode construite en ce point correspondant à $l = 1$. La méthode de réduction totale peut être interprétée comme une méthode à $l = \log_2 N_2$.

Les données du tableau 4 montrent qu'il existe, sous l'angle du coût en opérations arithmétiques, une méthode généralisée optimale avec $1 \leqslant l < \log_2 N_2$. L'analyse des estimations du nombre d'opérations dans la méthode à l réductions fournit une valeur optimale de $l = 1$ ou $l = 2$. Dans ce cas l'avantage minime gagné en nombre d'opérations par la méthode avec $l = 2$ peut disparaître du fait de la complicité accrue de l'algorithme.

2. Résolution des problèmes aux limites pour l'équation de Poisson dans un rectangle. Examinons maintenant comment on utilise la méthode construite au point 1 pour obtenir la solution de problèmes aux limites pour l'équation de Poisson dans un rectangle. Supposons qu'il s'agit de trouver dans le domaine $\bar{G} = \{0 \leqslant x_\alpha \leqslant \leqslant l_\alpha,\ \alpha = 1, 2\}$ la solution de l'équation

$$\frac{\partial^2 v}{\partial x_1^2} + \frac{\partial^2 v}{\partial x_2^2} = -\varphi(x), \quad x \in G, \tag{29}$$

vérifiant à la frontière Γ du rectangle $\bar{G}$ les conditions aux limites suivantes:

$$\begin{aligned} &\frac{\partial v}{\partial x_1} = \varkappa_{-1} v - g_{-1}(x_2), && x_1 = 0, \\ -&\frac{\partial v}{\partial x_1} = \varkappa_{+1} v - g_{+1}(x_2), && x_1 = l_1, \quad 0 \leqslant x_2 \leqslant l_2, \\ &\frac{\partial v}{\partial x_2} = -g_{-2}(x_1), && x_2 = 0, \\ -&\frac{\partial v}{\partial x_2} = -g_{+2}(x_1), && x_2 = l_2, \quad 0 \leqslant x_1 \leqslant l_1, \end{aligned} \tag{30}$$

où $\varkappa_{+1} \geqslant 0$, $\varkappa_{-1} \geqslant 0$, $\varkappa_{+1}^2 + \varkappa_{-1}^2 > 0$.

Admettons que dans les conditions (30) $\varkappa_{-1}$ et $\varkappa_{+1}$ sont des constantes. Avec cette hypothèse les inconnues dans le problème (29), (30) se séparent.

Sur le maillage rectangulaire $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G},\ 0 \leqslant \leqslant i \leqslant N_1,\ 0 \leqslant j \leqslant N_2,\ h_\alpha N_\alpha = l_\alpha,\ \alpha = 1, 2\}$ au problème (29)-(30) correspond le schéma aux différences

$$\Lambda u = (\Lambda_1 + \Lambda_2) u = -f(x), \quad x \in \bar{\omega}, \tag{31}$$

où $f(x) = \varphi(x) + \varphi_1(x) + \varphi_2(x)$,

$$\Lambda_1 u = \begin{cases} \frac{2}{h_1}(u_{x_1} - \varkappa_{-1} u), & x_1 = 0, \\ u_{\bar{x}_1 x_1}, & h_1 \leqslant x_1 \leqslant l_1 - h_1, \\ \frac{2}{h_1}(-u_{\bar{x}_1} - \varkappa_{+1} u), & x_1 = l_1; \end{cases}$$

$$\Lambda_2 u = \begin{cases} \frac{2}{h_2} u_{x_2}, & x_2 = 0, \\ u_{\bar{x}_2 x_2}, & h_2 \leqslant x_2 \leqslant l_2 - h_2, \\ -\frac{2}{h_2} u_{\bar{x}_2}, & x_2 = l_2, \end{cases}$$

quant aux fonctions $\varphi_\alpha(x)$, elles se déterminent par les relations

$$\varphi_\alpha(x) = \begin{cases} \frac{2}{h_\alpha} g_{-\alpha}(x_\beta), & x_\alpha = 0, \\ 0, & h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \quad \beta = 3 - \alpha, \quad \alpha = 1, 2, \\ \frac{2}{h_\alpha} g_{+\alpha}(x_\beta), & x_\alpha = l_\alpha. \end{cases}$$

On a montré dans le chapitre III que le schéma (31) présente sous la forme vectorielle l'écriture suivante:

$$\begin{aligned} &CU_0 - 2U_1 = F_0, \\ &-U_{j-1} + CU_j - U_{j+1} = F_j, \quad 1 \leqslant j \leqslant N_2 - 1, \\ &-2U_{N_2-1} + CU_{N_2} = F_{N_2}, \end{aligned} \qquad (32)$$

où

$$U_j = (u(0, j), u(1, j), \ldots, u(N_1, j)),$$
$$F_j = (h_2^2 f(0, j), h_2^2 f(1, j), \ldots, h_2^2 f(N_1, j)),$$
$$CU_j = ((2E - h_2^2 \Lambda_1) u(0, j), \ldots, (2E - h_2^2 \Lambda_1) u(N_1, j)),$$
$$0 \leqslant j \leqslant N_2.$$

Le système vectoriel (32) diffère du système (2) étudié auparavant par les conditions aux limites et la définition de la matrice C. Néanmoins, on construit sans peine l'analogue de la méthode du point 1 pour le problème (32). Puisque la déduction des principales formules de cette méthode ne diffère qu'en détails de celle exposée au point 2, on se limitera aux formules intermédiaires principales et finales. Pour la méthode de réduction totale les formules nécessaires sont décrites au § 4, ch. III.

Bref, pour les vecteurs $V_j = U_{2j}$, $0 \leqslant j \leqslant M_2$, où $2M_2 = N_2$, après l'élimination on aboutit au problème

$$\begin{aligned} &C^{(1)} V_0 - 2V_1 = \Phi_0, \\ &-V_{j-1} + C^{(1)} V_j - V_{j+1} = \Phi_j, \quad 1 \leqslant j \leqslant M_2 - 1, \\ &-2V_{M_2-1} + C^{(1)} V_{M_2} = \Phi_{M_2}, \end{aligned} \qquad (33)$$

où le second membre $\Phi_j = F_{2j}^{(1)}$, $0 \leqslant j \leqslant M_2$ se détermine suivant les formules

$$\Phi_j = \begin{cases} CF_0 + 2F_1, & j = 0, \\ F_{2j-1} + CF_{2j} + F_{2j+1}, & 1 \leqslant j \leqslant M_2 - 1, \\ CF_{N_2} + 2F_{N_2-1}, & j = M_2. \end{cases}$$

Pour les vecteurs V_j et Φ_j on a les développements

$$V_j = \sum_{k_2=0}^{M_2} Y_{k_2} \mu_{k_2}^{(2)}(j), \quad \Phi_j = \sum_{k_2=0}^{M_2} h_2^2 Z_{k_2} \mu_{k_2}^{(2)}(j), \quad 0 \leqslant j \leqslant M_2,$$

où

$$\mu_{k_2}^{(2)}(j) = \begin{cases} \dfrac{2}{\sqrt{l_2}} \cos \dfrac{k_2 \pi j}{M_2}, & 1 \leqslant k_2 \leqslant M_2 - 1, \\ \sqrt{\dfrac{1}{l_2}} \cos \dfrac{k_2 \pi j}{M_2}, & k_2 = 0, \ M_2. \end{cases}$$

Les coefficients de Fourier des vecteurs V_j et Φ_j en vertu de (33) sont liés par la relation

$$\left(C^{(1)} - 2 \cos \frac{k_2 \pi}{M_2} E \right) Y_{k_2} = h_2^2 Z_{k_2}, \quad 0 \leqslant k_2 \leqslant M_2,$$

tandis que les composantes du vecteur Z_{k_2} s'expriment au moyen des composantes du vecteur Φ_j de la façon suivante:

$$z_{k_2}(i) = \sum_{j=1}^{M_2-1} h_2 \varphi(i, j) \mu_{k_2}^{(2)}(j) + 0{,}5 h_2 [\varphi(i, 0) \mu_{k_2}^{(2)}(0) + \\ + \varphi(i, M_2) \mu_{k_2}^{(2)}(M_2)], \quad 0 \leqslant i \leqslant N_1.$$

Les inconnues U_j aux numéros j impairs, comme auparavant, se déterminent à partir des formules (4).

Il ne reste qu'à passer dans les formules obtenues à l'écriture scalaire et à la fonction propre non normalisée $\bar{\mu}_{k_2}^{(2)}(j) = \cos \frac{k_2 \pi j}{M_2}$.

On obtient finalement les formules suivantes pour la méthode de résolution du problème (31): pour chaque $0 \leqslant i \leqslant N_1$ on calcule

$$\varphi(i, j) = \begin{cases} 2[f(i, 0) + f(i, 1)] - h_2^2 \Lambda_1 f(i, 0), & j = 0, \\ f(i, 2j-1) + f(i, 2j+1) + 2f(i, 2j) - h_2^2 \Lambda_1 f(i, 2j), & \\ & 1 \leqslant j \leqslant M_2 - 1, \\ 2[f(i, N_2) + f(i, N_2 - 1)] - h_2^2 \Lambda_1 f(i, N_2), & j = M_2, \end{cases}$$

et on résout les équations

$$4 \sin^2 \frac{k_2 \pi}{2N_2} w_{k_2}(i) - h_2^2 \Lambda_1 w_{k_2}(i) = h_2^2 z_{k_2}(i), \quad 0 \leqslant i \leqslant N_1,$$

$$4 \cos^2 \frac{k_2 \pi}{2N_2} y_{k_2}(i) - h_2^2 \Lambda_1 y_{k_2}(i) = w_{k_2}(i), \quad 0 \leqslant i \leqslant N_1$$

pour $0 \leqslant k_2 \leqslant M_2$, où

$$z_{k_2}(i) = \sum_{j=0}^{M_2} \rho_j \varphi(i, j) \cos \frac{k_2 \pi j}{M_2},$$
$$0 \leqslant k_2 \leqslant M_2, \quad 0 \leqslant i \leqslant N_1.$$

La solution $u(i, j)$ du problème (31) s'obtient suivant les formules

$$u(i, 2j) = \sum_{k_2=0}^{M_2} \rho_{k_2} y_{k_2}(i) \cos \frac{k_2 \pi j}{M_2}, \quad 0 \leqslant j \leqslant M_2, \quad 0 \leqslant i \leqslant N_1$$

et à partir des équations

$$2u(i, 2j-1) - h_2^2 \Lambda_1 u(i, 2j-1) = \\ = h_2^2 f(i, 2j-1) + u(i, 2j-2) + u(i, 2j),$$
$$1 \leqslant j \leqslant M_2, \quad 0 \leqslant i \leqslant N_1.$$

On a utilisé ici les notations

$$\rho_j = \begin{cases} 1, & 1 \leqslant j \leqslant M_2 - 1, \\ 0{,}5, & j = 0,\ M_2, \quad M_2 = 0{,}5 N_2, \end{cases}$$

quant à l'opérateur Λ_1, il est déterminé plus haut. Pour trouver $w_{k_2}(i)$, $y_{k_2}(i)$ et $u(i, 2j-1)$, on dispose des équations triponctuelles aux conditions aux limites de troisième espèce qu'on résout à l'aide de la méthode du balayage.

Notons que les formules fournies ne changent nullement au cas où le maillage en direction de x_1 est irrégulier. Seul l'opérateur Λ_1 se modifie, ce sera l'analogue au sens de différences finies de la dérivée seconde et aux conditions aux limites de troisième espèce sur un maillage irrégulier.

En général il faut noter qu'il est possible de construire la variante adéquate de la méthode de séparation des variables avec estimation du coût du nombre d'opérations $O(N^2 \log_2 N)$ dans tous les cas, où il est possible d'utiliser la méthode de réduction totale, sauf un. L'exception concerne le cas où des conditions aux limites de troisième espèce sont imposées suivant la direction de l'élimination des inconnues au moins sur l'un des côtés du rectangle.

3. Problème de différences de Dirichlet d'ordre de précision élevé dans un rectangle. Examinons encore un exemple d'application de la méthode de séparation des variables. Etant donné un maillage rectangulaire $\bar{\omega}$, chercher la solution du *problème de différences de Dirichlet d'ordre de précision élevé pour l'équation de Poisson*

$$\Lambda u \left(\Lambda_1 + \Lambda_2 + \frac{h_1^2 + h_2^2}{12} \Lambda_1 \Lambda_2 \right) u = -f(x), \quad x \in \omega,$$
$$u(x) = 0, \quad x \in \gamma, \tag{34}$$

où $\Lambda_\alpha u = u_{\bar{x}_\alpha x_\alpha}$, $\alpha = 1, 2$. La condition aux limites pour simplifier est donnée homogène, le problème à condition aux limites inhomogène se réduit à (34) par correction du second membre de l'équation aux nœuds adjacents à la frontière.

Au point 4, § 1, ch. III on a obtenu la transcription vectorielle du problème (34) sous la forme suivante

$$-BU_{j-1} + AU_j - BU_{j+1} = F_j, \quad 1 \leqslant j \leqslant N_2 - 1,$$
$$U_0 = U_{N_2} = 0, \tag{35}$$

où

$$U_j = (u(1, j), u(2, j), \ldots, u(N_1 - 1, j)), \qquad 0 \leqslant j \leqslant N_2,$$
$$F_j = (h_2^2 f(1, j), h_2^2 f(2, j), \ldots, h_2^2 f(N_1 - 1, j)), \quad 1 \leqslant j \leqslant N_2 - 1,$$

quant aux matrices B et A, elles s'obtiennent à partir des relations

$$BU_j = \left(\left(E + \frac{h_1^2 + h_2^2}{12}\Lambda_1\right) u(1, j), \ldots \right.$$
$$\left. \ldots, \left(E + \frac{h_1^2 + h_2^2}{12}\Lambda_1\right) u(N_1 - 1, j)\right),$$
$$AU_j = \left(\left(2E - \frac{5h_2^2 - h_1^2}{6}\Lambda_1\right) u(1, j), \ldots \right.$$
$$\left. \ldots, \left(2E - \frac{5h_2^2 - h_1^2}{6}\Lambda_1\right) u(N_1 - 1, j)\right).$$

Les matrices A et B sont permutables, c'est-à-dire $AB = BA$.

Construisons la méthode combinée de séparation des variables pour le problème (34). D'abord effectuons la première élimination de la méthode de réduction pour le système (35). Faisons de ce pas une description indépendante de celle donnée au chapitre III. Ecrivons successivement trois équations du système (35) pour $j = 2, 4, 6, \ldots, N_2 - 2$:

$$-BU_{j-2} + AU_{j-1} - BU_j = F_{j-1},$$
$$-BU_{j-1} + AU_j - BU_{j+1} = F_j,$$
$$-BU_j + AU_{j+1} - BU_{j+2} = F_{j+1},$$

multiplions à gauche la première et la troisième équations par B, tandis que celle du milieu par A et additionnons-les. En vertu de la permutabilité de A et B, il vient

$$-B^2U_{j-2} + (A^2 - 2B^2)U_j - B^2U_{j+2} = F_j^{(1)},$$
$$j = 2, 4, 6, \ldots, N_2 - 2,$$
$$U_0 = U_N = 0,$$

où $F_j^{(1)} = B(F_{j-1} + F_{j+1}) + AF_j$, $j = 2, 4, 6, \ldots, N_2 - 2$. Posons, comme habituellement, $V_j = U_{2j}$, $0 \leqslant j \leqslant M_2$, $\Phi_j = F_{2j}^{(1)}$, $1 \leqslant j \leqslant M_2 - 1$, où $2M_2 = N_2$ et écrivons le système sous la

forme

$$-B^2V_{j-1}+(A^2-2B^2)V_j-B^2V_{j+1}=\Phi_j,\quad 1\leqslant j\leqslant M_2-1,$$
$$V_0=V_{M_2}=0, \tag{36}$$

en outre

$$\Phi_j=B(F_{2j-1}+F_{2j+1})+AF_{2j},\quad 1\leqslant j<M_2-1. \tag{37}$$

Les vecteurs inconnus restants se déterminent à partir des équations

$$AU_{2j-1}=F_{2j-1}+B(U_{2j-2}+U_{2j}),\quad 1\leqslant j\leqslant M_2. \tag{38}$$

Le système « raccourci » (36) sera résolu, comme auparavant, par la méthode de Fourier. Portons les développements (12) dans (36), où $\mu_{k_2}^{(2)}(j)$ sont définis dans (10). Finalement, pour les coefficients de Fourier Y_{k_2} et Z_{k_2} des vecteurs V_j et Φ_j on obtient la relation

$$\left(A^2-4\cos^2\frac{k_2\pi}{2M_2}B^2\right)Y_{k_2}=h_2^2Z_{k_2},\quad 1\leqslant k_2\leqslant M_2-1, \tag{39}$$

qui est l'analogue de la relation (13), les composantes des vecteurs Z_{k_2} et Φ_j étant liées par la formule (11). Pour résoudre l'équation (39) on peut profiter de l'algorithme

$$\left(A-2\cos\frac{k_2\pi}{2M_2}B\right)W_{k_2}=h_2^2Z_{k_2},$$
$$\left(A+2\cos\frac{k_2\pi}{2M_2}B\right)Y_{k_2}=W_{k_2},\quad 1\leqslant k\leqslant M_2-1. \tag{40}$$

Bref, la méthode de résolution du problème (34) se décrit en forme vectorielle par les formules (37), (11), (40), (12) et (38). En passant à l'écriture scalaire et à la fonction propre non normée $\bar\mu_{k_2}^{(2)}(j)=\sin\frac{k_2\pi j}{M_2}$ au moyen de la substitution tirée du point 1, on obtient les formules suivantes:

$$\varphi(i,j)=\left(E+\frac{h_1^2+h_2^2}{12}\Lambda_1\right)[f(i,2j-1)+f(i,2j+1)+2f(i,2j)]-$$
$$-h_2^2\Lambda_1 f(i,2j),\quad 1\leqslant j\leqslant M_2-1,\quad 1\leqslant i\leqslant N_1-1, \tag{41}$$
$$f(0,j)=0,\quad 1\leqslant j\leqslant N_1-1$$

permettant de calculer $\varphi(i,j)$: les équations

$$4\sin^2\frac{k_2\pi}{2N_2}w_{k_2}(i)-h_2^2\left(1-\frac{4}{h_2^2}\sin^2\frac{k_2\pi}{2N_2}\cdot\frac{h_1^2+h_2^2}{12}\right)\Lambda_1 w_{k_2}(i)=$$
$$=h_2^2z_{k_2}(i), \tag{42}$$
$$1\leqslant i\leqslant N_1-1,\quad w_{k_2}(0)=w_{k_2}(N_1)=0$$

pour le calcul de $w_{k_2}(i)$ et

$$4\cos^2\frac{k_2\pi}{2N_2}y_{k_2}(i)-h_2^2\left(1-\frac{4}{h_2^2}\cos^2\frac{k_2\pi}{2N_2}\,\frac{h_1^2+h_2^2}{12}\right)\Lambda_1 y_{k_2}(i)=w_{k_2}(i), \tag{43}$$
$$1\leqslant i\leqslant N_1-1,\quad y_{k_2}(0)=y_{k_2}(N_2)=0$$

pour le calcul de $y_{k_2}(i)$ qu'on peut résoudre pour $1 \leqslant k_2 \leqslant$ $\leqslant M_2-1$, où

$$z_{k_2}(i)=\sum_{j=1}^{M_2-1} \varphi(i,\ j) \sin \frac{k_2 \pi j}{M_2}, \quad 1 \leqslant k_2 \leqslant M_2-1, \quad 1 \leqslant i \leqslant N_1-1. \tag{44}$$

La solution $u(i,\ j)$ du problème (34) s'obtient suivant les formules

$$u(i,\ 2j)=\frac{4}{N_2} \sum_{k_2=1}^{M_2-1} y_{k_2}(i) \sin \frac{k_2 \pi j}{M_2}, \quad 1 \leqslant j \leqslant M_2-1, \quad 1 \leqslant i \leqslant N_1-1, \tag{45}$$

et à partir des équations

$$\begin{gathered}2u(i,\ 2j-1)-\frac{5h_2^2-h_1^2}{6} \Lambda_1 u(i,\ 2j-1)=h_2^2 f(i,\ 2j-1)+ \\ +\left(E+\frac{h_1^2+h_2^2}{12} \Lambda_1\right)[u(i,\ 2j-2)+u(i,\ 2j)], \\ 1 \leqslant i \leqslant N_1-1, \\ u(0,\ 2j-1)=u(N_1,\ 2j-1)=0, \quad 1 \leqslant j \leqslant M_2.\end{gathered} \tag{46}$$

Il nous reste à montrer que les équations triponctuelles (42), (43) et (46) admettent une solution. On peut alors, pour obtenir la solution, utiliser la méthode triviale du balayage ou la méthode du balayage non monotone.

Il suffit de montrer que pour $1 \leqslant k_2 \leqslant N_2-1$ les valeurs propres de l'opérateur de différences

$$\mathscr{R}=\lambda_{k_2}^{(2)} E-\left(1-\frac{h_1^2+h_2^2}{12} \lambda_{k_2}^{(2)}\right) \Lambda_1, \quad \lambda_{k_2}^{(2)}=\frac{4}{h_2^2} \sin^2 \frac{k_2 \pi}{2N_2}$$

sont différentes de zéro. En effet, pour $1 \leqslant k_2 \leqslant N_2/2-1$ l'opérateur $h_2^2 \mathscr{R}$ coïncide avec l'opérateur du problème (42), tandis que pour $k_2=N_2/2$ il coïncide avec l'opérateur du problème (46). Si $N_2/2+1 \leqslant k_2 \leqslant N_2-1$ l'opérateur $h_2^2 \mathscr{R}$ prend la forme

$$h_2^2 \mathscr{R}=4 \sin^2 \frac{k_2 \pi}{2N_2}-h_2^2\left(1-\frac{h_1^2+h_2^2}{12} \frac{4}{h_2^2} \sin^2 \frac{k_2 \pi}{2N_2}\right) \Lambda_1.$$

La substitution $k_2=N_2-k_2'$ donne

$$h_2^2 \mathscr{R}=4 \cos^2 \frac{k_2' \pi}{2N_2}-h_2^2\left(1-\frac{h_1^2+h_2^2}{12} \frac{4}{h_2^2} \cos^2 \frac{k_2' \pi}{2N_2}\right) \Lambda_1,$$

où $1 \leqslant k_2' \leqslant N_2/2-1$, c'est-à-dire que dans ce cas l'opérateur $h_2^2 \mathscr{R}$ coïncide avec celui du problème (43).

Cherchons maintenant les valeurs propres de l'opérateur $\mathscr{R}$ pour un k_2 fixé. Comme les valeurs propres de l'opérateur Λ_1 au cas des

conditions aux limites de première espèce sont (voir § 5, ch. I)

$$\lambda_{k_1}^{(1)}=\frac{4}{h_1^2}\sin^2\frac{k_1\pi}{2N_1}, \quad k_1=1, 2, \ldots, N_1-1,$$

les valeurs propres λ de l'opérateur $\mathscr{R}$ sont

$$\lambda_{k_1k_2}=\lambda_{k_1}^{(1)}+\lambda_{k_2}^{(2)}-\frac{h_1^2+h_2^2}{12}\lambda_{k_1}^{(1)}\lambda_{k_2}^{(2)}, \quad 1\leqslant k_1\leqslant N_1-1, \ 1\leqslant k_2\leqslant N_2-1.$$

Etant donné qu'on a les estimations suivantes des valeurs propres $\lambda_{k_1}^{(1)}$ et $\lambda_{k_2}^{(2)}$:

$$0<\lambda_{k_\alpha}^{(\alpha)}<\frac{4}{h_\alpha^2}, \quad \alpha=1, 2,$$

il est possible d'obtenir sans peine pour k_1 et k_2 quelconques

$$\lambda_{k_1k_2}=\lambda_{k_1}^{(1)}\left(1-\frac{h_2^2}{12}\lambda_{k_2}^{(2)}\right)+\lambda_{k_2}^{(2)}\left(1-\frac{h_1^2}{12}\lambda_{k_1}^{(1)}\right)>\frac{2}{3}(\lambda_{k_1}^{(1)}+\lambda_{k_2}^{(2)})>0,$$

ce qu'il fallait démontrer.

On trouve aisément que pour le problème (42) la condition suffisante de l'applicabilité de la méthode du balayage trivial prend la forme

$$1+\frac{2h_1^2-h_2^2}{3h_2^2}\sin^2\frac{k_2\pi}{2N_2}\geqslant 0 \qquad (47)$$

et qu'elle est évidemment vérifiée pour tout k_2. Pour (43) la condition analogue est de la forme

$$1+\frac{2h_1^2-h_2^2}{3h_2^2}\cos^2\frac{k_2\pi}{2N_2}\geqslant 0$$

qui est également vraie pour tous les k_2. Au problème (46) correspond la condition (47) avec $k_2 = 0{,}5N_2$. Par conséquent, les problèmes (42), (43) et (46) se prêtent à la résolution par la méthode du balayage trivial.

CHAPITRE V

APPAREIL MATHÉMATIQUE DE LA THÉORIE DES MÉTHODES ITÉRATIVES

Ce chapitre fournit des renseignements ainsi que des notions principales sur la théorie des méthodes itératives exposées dans les chapitres suivants. Au § 1 sont données les notions les plus simples de l'analyse fonctionnelle et fournies les principales propriétés des opérateurs linéaires et non linéaires dans l'espace hilbertien, ainsi que quelques théorèmes sur la résolubilité des équations opératorielles. Le § 2 est consacré à l'interprétation systématique des schémas aux différences comme des équations opératorielles dans un espace abstrait avec indication des propriétés des opérateurs associés. Au § 3 sont présentées les principales définitions et notions de la théorie des procédés itératifs, est examinée la forme canonique des schémas itératifs, sont fournies des notions sur la convergence et le nombre d'itérations.

§ 1. Eléments d'information sur l'analyse fonctionnelle

1. Espaces linéaires. Dans les chapitres précédents on a étudié les principales méthodes directes de résolution des équations aux différences les plus simples. Les méthodes élaborées se caractérisent par le fait qu'avec leur aide il est en principe possible, en réalisant un certain nombre fini d'opérations, d'obtenir une solution précise du problème de différences. Il est naturellement admis dans ce cas que l'information d'entrée est précise et que les calculs sont conduits sans arrondi.

L'efficience des ces méthodes est suffisamment élevée, vu la prise en compte de la structure matricielle du système résolu. L'obligation de se plier à certaines propriétés spéciales des matrices rétrécit le champ d'applicabilité de ces méthodes en le limitant aux problèmes les plus simples.

Pour la résolution de problèmes compliqués et, en particulier, de problèmes de différences non linéaires on utilise habituellement des méthodes itératives. Le principe des méthodes itératives réside dans la construction par un mode quelconque d'approximations successives aboutissant à la solution, en commençant par une certaine approximation initiale. Pour solution approchée du problème, on adopte dans ce cas la solution obtenue après un nombre fini d'itérations.

L'universalité des méthodes itératives réside avant tout dans le fait qu'elles permettent de résoudre non pas un problème concret mais une classe de problèmes possédant des propriétés déterminées. Ces propriétés ne sont pas fonction de la structure des équations de mailles mais des propriétés fonctionnelles générales. Vu que dans la plupart des méthodes itératives on néglige la structure concrète des équations, on est en mesure de construire la théorie des méthodes itératives sous une optique unique en concentrant l'étude sur l'équation opératorielle de première espèce

$$Au = f,$$

où A est l'opérateur, f l'élément donné et u l'élément cherché d'un certain espace H.

Avant de passer à la construction et à l'étude des méthodes itératives, donnons une information sommaire sur l'analyse fonctionnelle (sans esquisser de démonstrations).

On appelle *espace linéaire* sur un champ K de nombres réels ou complexes l'ensemble H pour les éléments duquel sont définies des opérations d'addition des éléments et de multiplication de l'élément par un nombre du champ K, avec vérification des axiomes suivants (x, y, z — éléments de H, λ et μ nombres de K):

1) les deux opérations n'entraînent pas la sortie de H;

2) $x + y = y + x$, $x + (y + z) = (x + y) + z$ (commutativité et associativité de l'addition);

3) $\lambda(\mu x) = (\lambda\mu) x$ (associativité de la multiplication);

4) $\lambda(x + y) = \lambda x + \lambda y$, $(\lambda + \mu) x = \lambda x + \mu x$ (distributivité de la multiplication relativement à l'addition);

5) il existe de façon univoque un certain élément 0 pour lequel $x + 0 = x$ pour tout $x \in H$;

6) il existe de façon univoque pour chaque $x \in H$ un élément $(-x) \in H$ pour lequel $x + (-x) = 0$;

7) $1 \cdot x = x$.

Suivant que les nombres par lesquels est tolérée la multiplication des éléments de H sont réels ou complexes, on aura un *espace linéaire réel* ou *complexe*.

On peut introduire dans les espaces linéaires la notion de dépendance et d'indépendance linéaires des éléments. Les éléments $x_1, x_2, \ldots, x_n$ de l'espace linéaire H sont *linéairement indépendants* si de l'égalité

$$\lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_n x_n = 0 \tag{1}$$

il s'ensuit que $\lambda_1 = \lambda_2 = \ldots = \lambda_n = 0$. Si, au contraire, il se trouve parmi les $\lambda_1, \lambda_2, \ldots, \lambda_n$ non tous nuls de tels pour lesquels (1) est vérifié, alors les éléments $x_1, x_2, \ldots, x_n$ sont appelés *linéairement dépendants*.

L'espace H est dit à *n dimensions* s'il existe dans H n éléments linéairement indépendants, tandis que tout $(n+1)$-ième élément est linéairement dépendant.

Un ensemble H_1 fermé non vide d'éléments de l'espace linéaire H est appelé *sous-espace* si à côté des éléments $x_1, x_2, \ldots, x_n$ l'ensemble H_1 comprend toute combinaison linéaire $\lambda_1 x_1 + \lambda_2 x_2 + \ldots$ $\ldots + \lambda_n x_n$ de ces éléments.

La somme du nombre fini des sous-espaces $H_1, H_2, \ldots, H_n$ constitue un ensemble d'éléments de la forme

$$x = x_1 + x_2 + \ldots + x_n, \quad x_i \in H_i, \ i = 1, 2, \ldots, n. \qquad (2)$$

Soient $H_1, H_2, \ldots, H_n$ les sous-espaces appartenant à l'espace linéaire H. Si chaque élément $x \in H$ se représente univoquement sous la forme (2), on dit alors que H est une *somme directe des sous-espaces* $H_1, H_2, \ldots, H_n$, quant à l'expression (2), elle est appelée *développement de l'élément x en éléments de* $H_1, H_2, \ldots, H_n$.

Dans ce cas il vient

$$H = H_1 \oplus H_2 \oplus \ldots \oplus H_n.$$

On montrera sans peine que si $H = H_1 \oplus H_2$, H_1 et H_2 n'ont comme élément commun que l'élément nul de l'espace. Inversement, si un élément quelconque $x \in H$ peut se représenter sous forme de $x = x_1 + x_2$, $x_1 \in H_1$, $x_2 \in H_2$ et $H_1 \cap H_2 = 0$, on a alors $H = H_1 \oplus H_2$.

L'espace linéaire H est dit *normé* si pour chaque élément $x \in H$ est défini un nombre réel $\| x \|$ appelé *norme* vérifiant les conditions:

1) $\| x \| \geqslant 0$, avec $\| x \| = 0$, si $x = 0$;
2) $\| x + y \| \leqslant \| x \| + \| y \|$ (inégalité du triangle);
3) $\| \lambda x \| = | \lambda | \, \| x \|$, λ étant un nombre.

La suite $\{x_n\}$ d'éléments de l'espace linéaire normé H est dite *convergente* vers l'élément $x \in H$ si $\| x - x_n \| \to 0$ pour $n \to \infty$. Si $\| x_n - x_m \| \to 0$ pour $n, m \to \infty$, la suite $\{x_n\}$ est dite *fondamentale* (*suite de Cauchy*).

L'espace linéaire normé H est dit *complet* si toute suite de Cauchy $\{x_n\}$ de cet espace est convergente vers un certain élément $x \in H$. Les espaces linéaires normés complets sont appelés *espaces de Banach*. Tout espace linéaire normé à dimensions finies est complet. Les sous-espaces de l'espace normé sont normés de façon naturelle.

Un même espace linéaire peut être normé de façon infinie. Soient les normes $\| x \|_1$ et $\| x \|_2$ imposées de deux façons différentes à un espace linéaire. S'il existe des constantes $0 < m \leqslant M$ qui pour tout $x \in H$ vérifient les inégalités

$$m \, \| x \|_1 \leqslant \| x \|_2 \leqslant M \, \| x \|_1,$$

ces normes sont alors dites *équivalentes*. Notons que dans un espace à dimensions finies toutes deux normes sont équivalentes.

Si dans un espace linéaire on a introduit deux normes équivalentes, la convergence d'une certaine suite $\{x_n\}$ dans l'une des normes implique la convergence dans l'autre.

Soit H un espace linéaire réel (complexe) et soit opposé à deux éléments x, y de H un nombre réel (complexe) (x, y) tel que

1) $(x, y) = \overline{(y, x)}$ (symétrie);
2) $(x + y, z) = (x, z) + (y, z)$ (distributivité);
3) $(\lambda x, y) = \lambda (x, y)$ (homogénéité);
4) $(x, x) \geqslant 0$ pour tout $x \in H$, avec $(x, x) = 0$ seulement et rien que seulement pour $x = 0$.

Le nombre (x, y) est appelé *produit scalaire* des éléments x et y. Le trait au-dessus signifie qu'il y a passage au nombre complexe conjugué.

L'espace linéaire normé H dans lequel la norme est introduite par le produit scalaire $\| x \| = \sqrt{(x, x)}$ est appelé espace *unitaire H*. L'espace unitaire complet est dénommé *espace de Hilbert* (ou hilbertien). L'espace unitaire à dimensions finies est complet.

Pour un produit scalaire se vérifie l'inégalité de Cauchy-Bouniakovski $|(x, y)| \leqslant \| x \| \, \| y \|$. Les éléments x et y de l'espace unitaire sont dits *mutuellement orthogonaux* si $(x, y) = 0$. L'élément $x \in H$ est appelé *sous-espace orthogonal à H_1 de l'espace H* si x est orthogonal à tout élément $y \in H_1$. L'ensemble H_2 de tous les éléments $x \in H$ orthogonaux au sous-espace H_1 de l'espace H est dénommé *complément orthogonal* du sous-espace H_1. Notons que le complément orthogonal constitue lui-même un sous-espace de l'espace H.

Soit H_1 un sous-espace quelconque de l'espace H et H_2 le complément orthogonal. Dans ce cas H est une somme directe de H_1 et H_2, $H = H_1 \oplus H_2$. Par conséquent, chaque élément $x \in H$ se représente de façon unique sous forme de $x = x_1 + x_2$, $x_\alpha \in H_\alpha$, $\alpha = 1, 2$, avec $(x_1, x_2) = 0$.

Le système $x_1, x_2, \ldots, x_n$ d'éléments de l'espace H est appelé *système orthogonal* si $(x_m, x_n) = \delta_{mn}$, $m, n = 1, 2, \ldots$, où δ_{mn} est le symbole de Kronecker (delta de Kronecker) qui est égal à l'unité pour $m = n$ et à zéro pour $m \neq n$.

S'il n'existe pas d'élément $x \in H$ différant de zéro et orthogonal à tous les éléments du système orthonormé $\{x_n\}$, ce système est dit *complet*. La série de Fourier $\sum_{k=1}^{\infty} c_k x_k$, où $c_k = (x, x_k)$, $k = 1, 2, \ldots$, construite pour tout $x \in H$ suivant le système orthonormé complet $\{x_n\}$, converge vers cet élément, et pour tout $x \in H$ on a l'égalité

$$\| x \|^2 = (x, x) = \sum_{k=1}^{\infty} c_k^2.$$

2. Opérateurs dans des espaces linéaires normés. Soient X et Y les espaces linéaires normés. On dit que sur l'ensemble $\mathscr{D} \subset X$

est donné l'opérateur A aux valeurs en Y (opérateur agissant de $\mathscr{D}$ en Y), si à chaque élément $x \in \mathscr{D}$ correspond l'élément $y = Ax \in Y$. L'ensemble $\mathscr{D}$ est appelé *domaine de définition de l'opérateur A* et est désigné par $\mathscr{D}(A)$. L'ensemble de tous les éléments $y \in Y$ représentés sous forme de $y = Ax$ $(x \in \mathscr{D}(A))$ est dénommé *domaine des valeurs de l'opérateur A* et est noté im A. Si $\mathscr{D}(A) = X$, im $A \subset X$, c'est-à-dire que l'opérateur A est une application de X en lui-même, on dit que A est un opérateur sur X.

L'opérateur A est dénommé *linéaire* quand $\mathscr{D}(A)$ est une multiplicité linéaire dans X et pour tous les $x_1, x_2 \in \mathscr{D}(A)$

$$A(\lambda_1 x_1 + \lambda_2 x_2) = \lambda_1 A x_1 + \lambda_2 A x_2,$$

où λ_1 et λ_2 sont des nombres du champ K.

L'opérateur linéaire A est dit *borné* s'il existe une constante $M > 0$ qui pour tous les $x \in \mathscr{D}(A)$ vérifie l'inégalité

$$\| Ax \|_2 \leqslant M \| x \|_1, \tag{3}$$

où $\|\cdot\|_1$ est la norme dans X, $\|\cdot\|_2$ la norme dans Y. L'opérateur non linéaire arbitraire A est dit *borné* sur $\mathscr{D}(A)$ si

$$\sup_{x \in \mathscr{D}(A)} \| Ax \|_2 < \infty.$$

On appelle *norme* de l'opérateur en la notant par $\| A \|$ la plus petite des constantes M vérifiant la condition (3) pour l'opérateur linéaire A. Il s'ensuit de la définition de la norme que

$$\| A \| = \sup_{\| x \|_1 = 1} \| Ax \|_2 \quad \text{ou} \quad \| A \| = \sup_{x \neq 0} \frac{\| Ax \|_2}{\| x \|_1}.$$

Remarquons que dans un espace à dimensions finies tout opérateur linéaire est borné. Soit A un opérateur quelconque agissant de X dans Y. L'opérateur A est dit *continu au point $x \in X$*, si de la condition $\| x_n - x \|_1 \to 0$ $(x_n \in X)$ il s'ensuit que $\| Ax_n - Ax \|_2 \to 0$ pour $n \to \infty$. L'opérateur linéaire borné est continu.

L'opérateur arbitraire A vérifie la *condition de Lipschitz à constante q* si

$$\| Ax_1 - Ax_2 \|_2 \leqslant q \| x_1 - x_2 \|_1, \quad x_1, x_2 \in \mathscr{D}(A). \tag{4}$$

Tout opérateur linéaire borné A satisfait à la condition de Lipschitz (4) avec $q = \| A \|$.

Soit A un opérateur quelconque agissant de X dans Y. L'opérateur linéaire borné $A'(x)$ est appelé *dérivée Gateau de l'opérateur A au point x* de l'espace X si pour tout $z \in X$ on a

$$\lim_{t \to 0} \left\| \frac{A(x+tz) - Ax}{t} - A'(x) z \right\|_2 = 0.$$

En outre, le domaine des valeurs de l'opérateur A' appartient à Y.

Si l'opérateur A possède une dérivée Gateau en chaque point de l'espace X, alors pour tous $x_1, x_2 \in X$ se vérifie l'inégalité (4),

où $q = \sup_{0 \leqslant t \leqslant 1} \| A' (x_1 + t (x_2 - x_1)) \|$. Si A est un opérateur linéaire, $A' = A$.

Tous les opérateurs linéaires bornés pensables qui agissent de X dans Y constituent un *espace linéaire normé*, car la norme $\| A \|$ de l'opérateur A vérifie toutes les axiomes de la norme. Examinons un ensemble d'opérateurs linéaires bornés agissant de X dans X. On peut introduire sur cet ensemble le produit AB d'opérateurs A et B de la façon suivante: $(AB)\, x = A\, (Bx)$. AB est apparemment un opérateur linéaire borné: $\| AB \| \leqslant \| A \| \, \| B \|$.

Si $(AB)\, x = (BA)\, x$ pour tous les $x \in X$, les opérateurs A et B s'appellent alors de *permutation* ou *commutatifs*; on écrit dans ce cas $AB = BA$.

En rapport avec la résolution des équations de la forme $Ax = y$ on a introduit la notion d'opérateur *inverse* A^{-1}. Soit A l'opérateur de X sur Y. Si à chaque $y \in Y$ ne correspond qu'un $x \in X$ pour lequel $Ax = y$, cette correspondance détermine alors l'opérateur A^{-1} appelé *inverse* de A et présentant un domaine de définition Y et un domaine de valeurs X.

Pour tous $x \in X$ et $y \in Y$ on a des identités $A^{-1} (Ax) = x$, $A\, (A^{-1} y) = y$. On montre sans peine que si A est linéaire, A^{-1} (s'il existe) est également linéaire.

Lemme 1. *Pour qu'un opérateur linéaire A constituant une application de X sur Y possède un opérateur inverse, il faut et il suffit que $Ax = 0$ rien que pour $x = 0$.*

Théorème 1. *Soit A un opérateur linéaire de X sur Y. Pour qu'un opérateur inverse A^{-1} existe et soit borné (comme l'est l'opérateur de Y sur X) il faut et il suffit qu'il existe une telle constante $\delta > 0$ pour laquelle pour tous les $x \in X$ on ait*

$$\| Ax \|_2 \geqslant \delta \, \| x \|_1 .$$

Dans ce cas se vérifie l'estimation $\| A^{-1} \| \leqslant 1/\delta$. $\| \cdot \|_1$ est ici la norme dans X, et $\| \cdot \|_2$ la norme dans Y.

En d'autres termes, pour que l'opérateur inverse A^{-1} existe il faut et il suffit que l'équation homogène $Ax = 0$ ne possède qu'une solution triviale.

Soient A et B les opérateurs linéaires bornés agissant dans X et possédant des inverses. Dans ce cas $(AB)^{-1} = B^{-1} A^{-1}$.

Si l'opérateur A est inversible, alors les puissances A^k aux exposants entiers quelconques (et non seulement négatifs) acquièrent un sens. Et, notamment, par définition $A^{-k} = (A^{-1})^k$, $k = 1, 2, \ldots$ Les puissances d'un même opérateur commutent.

Introduisons la notion de *noyau* de l'opérateur linéaire A. On appelle *noyau de l'opérateur linéaire A* l'ensemble de tous les élé-

ments x de l'espace X pour lesquels $Ax = 0$. Le noyau de l'opérateur linéaire A est désigné par le symbole $\ker A$.

La condition $\ker A = 0$ est nécessaire et suffisante pour que l'opérateur A possède un inverse.

Le sous-espace X_1 de l'espace X est dit sous-espace *invariant* de l'opérateur A agissant dans X si A n'implique pas la sortie des éléments de X_1, c'est-à-dire $Ax \in X_1$ si $x \in X_1$.

Si le sous-espace X_1 est invariant relativement à l'opérateur inversible A, il est invariant relativement à l'opérateur A^{-1}.

En guise d'exemples de sous-espaces invariants de l'opérateur A citons $\ker A$ et $\operatorname{im} A$. Notons que si les opérateurs A et B commutent, les sous-espaces $\ker B$ et $\operatorname{im} B$ sont invariants relativement à l'opérateur A.

Le nombre

$$\rho(A) = \lim_{k \to \infty} \sqrt[k]{\| A^k \|}$$

est appelé *rayon spectral de l'opérateur linéaire A*. Il ne dépend pas de la définition de la norme, de plus $\rho(A) = \inf_{\|\cdot\|} \| A \|$.

Pour tout opérateur linéaire borné A se vérifient les inégalités

$$\rho(A) \leqslant \| A \|, \quad \rho(A) \leqslant \sqrt[k]{\| A^k \|}, \quad k = 2, 3, \ldots$$

L e m m e 2. *Pour que* $\| A \| = \rho(A)$, *il faut et il suffit que* $\| A^k \| = \| A \|^k$, $k = 2, 3, \ldots$

Notons encore une propriété du rayon spectral. Si les opérateurs A et B commutent, on a alors

$$\rho(AB) \leqslant \rho(A)\,\rho(B), \qquad \rho(A + B) \leqslant \rho(A) + \rho(B).$$

3. Opérateurs dans l'espace hilbertien. Soit un opérateur linéaire borné A agissant dans l'espace unitaire H. En conformité avec la définition générale de la norme de l'opérateur, il vient

$$\| A \| = \sup_{\| x \| = 1} \| Ax \| = \sup_{x \in H} \sqrt{\frac{(Ax, Ax)}{(x, x)}}$$

et, par conséquent, pour tout $x \in H$ se vérifie l'inégalité

$$(Ax, Ax) \leqslant \| A \|^2 (x, x).$$

En utilisant l'inégalité de Cauchy-Bouniakovski, on en obtient

$$| (Ax, x) | \leqslant \| Ax \| \, \| x \| \leqslant \| A \| (x, x). \tag{5}$$

On ne c o n s i d é r e r a d a n s l a s u i t e q u e d e s o p é r a t e u r s b o r n é s.

L'opérateur A^* est appelé *adjoint* (conjugué) de l'opérateur A si pour tous x, $y \in H$ est vérifiée l'identité

$$(Ax, y) = (x, A^*y).$$

Pour tout opérateur linéaire borné A avec le domaine de définition $\mathscr{D}(A) = H$ il existe un opérateur A^*, qui d'ailleurs est unique, avec le domaine de définition $\mathscr{D}(A^*) = H$. L'opérateur A^* est linéaire et borné, $\|A^*\| = \|A\|$.

Formulons les principales propriétés de l'opération de conjugaison: $(A^*)^* = A$, $(A + B)^* = A^* + B^*$, $(AB)^* = B^*A^*$, $(\lambda A)^* = \bar{\lambda} A^*$. Si les opérateurs A et B commutent, les opérateurs adjoints A^* et B^* commutent également. Si A possède un inverse, $(A^{-1})^* = (A^*)^{-1}$, c'est-à-dire que l'inversion et la conjugaison de l'opérateur sont des opérations permutables.

L e m m e 3. *Soit A un opérateur linéaire dans H. L'espace H peut être représenté sous forme de sommes directes des sous-espaces orthogonaux*

$$H = \ker A \oplus \operatorname{im} A^*, \quad H = \ker A^* \oplus \operatorname{im} A.$$

En effet, soit H_1 le complément orthogonal im A^* jusqu'à l'espace H, c'est-à-dire

$$H = H_1 \oplus \operatorname{im} A^*, \ (x_1, x_2) = 0, \ x_1 \in H_1, \ x_2 \in \operatorname{im} A^*.$$

Montrons que $H_1 = \ker A$. Soit $x_1 \in \ker A$, dans ce cas pour tou- $x \in H$ on a $A^*x \in \operatorname{im} A^*$ et

$$(x_1, A^*x) = (Ax_1, x) = 0.$$

Donc x_1 est orthogonal à im A^* et, partant, $x_1 \in H_1$. D'autre part, soit $x_1 \in H_1$ (donc x_1 est orthogonal à im A^*). Alors pour tout $x \in H$

$$0 = (x_1, A^*x) = (Ax_1, x).$$

Vu que x est un élément quelconque de H, $Ax_1 = 0$ et, partant, $x_1 \in \ker A$. La première proposition du lemme est démontrée. De façon analogue on démontre la seconde proposition.

L'opérateur linéaire A est appelé *autoadjoint* (autoconjugué) dans H si $A = A^*$. Pour un opérateur autoadjoint $(Ax, y) = (x, Ay)$, quels que soient x, $y \in H$.

L'opérateur A est dit *normal* s'il commute avec son adjoint, $A^*A = AA^*$ et il est dit de *symétrie gauche* si $A^* = -A$. Les opérateurs autoadjoints et de symétrie gauche sont normaux.

Comme on sait, si A et B sont des opérateurs autoadjoints l'opérateur AB est autoadjoint seulement et rien que seulement si A et B sont permutables.

Si A est un opérateur linéaire, A^*A et AA^* sont autoadjoints, de plus $\| A^*A \| = \| AA^* \| = \| A \|^2$ et

$$\ker A^*A = \ker A, \quad \operatorname{im} A^*A = \operatorname{im} A^*,$$

$$\ker AA^* = \ker A^*, \quad \operatorname{im} AA^* = \operatorname{im} A.$$

Tout opérateur A peut être représenté sous forme de somme d'opérateurs autoadjoint A_0 et de symétrie gauche A_1

$$A = A_0 + A_1,$$

où $A_0 = 0{,}5\,(A + A^*)$, $A_1 = 0{,}5\,(A - A^*)$. Si H est un espace réel, il s'ensuit les égalités

$$(Ax, x) = (A_0x, x), \quad (A_1x, x) = 0.$$

Dans un espace complexe H on a une représentation cartésienne de l'opérateur A :

$$A = A_0 + iA_1,$$

où $A_0 = \operatorname{Re} A = \frac{1}{2}(A + A^*)$, $A_1 = \operatorname{Im} A = \frac{1}{2i}(A - A^*)$ sont des opérateurs autoadjoints dans H. De plus, pour tous $x \in H$ se vérifient les identités

$$\operatorname{Re}(Ax, x) = (A_0x, x), \quad \operatorname{Im}(Ax, x) = (A_1x, x).$$

Si A est un opérateur autoadjoint dans H, on a la formule

$$\| A \| = \sup_{x \neq 0} \frac{|(Ax, x)|}{(x, x)}, \quad x \in H.$$

L e m m e 4. *Si A est un opérateur autoadjoint borné dans H, alors pour tout n entier plus grand que zéro se vérifie l'égalité* $\| A^n \| = \| A \|^n$.

Le lemme 4 reste vrai également pour un opérateur normal.

Il s'ensuit des lemmes 2 et 4 que pour un opérateur A normal (en particulier, autoadjoint) on a l'égalité $\rho(A) = \| A \|$.

L e m m e 5. *Soit dans un espace linéaire H introduit au moyen de deux procédés le produit scalaire des éléments x et y: $(x, y)_1$ et $(x, y)_2$. Si l'opérateur A est autoadjoint au sens de chaque produit scalaire, on a alors* $\| A \|_1 = \| A \|_2 = \rho(A)$.

Le rayon spectral fournit une estimation par le bas pour toute norme de l'opérateur. Introduisons le rayon numérique de l'opérateur permettant d'obtenir les estimations de la norme dans les deux sens.

Le *rayon numérique de l'opérateur A* agissant dans un espace complexe H peut être défini de la façon suivante:

$$\bar{\rho}(A) = \sup_{\| x \| = 1} |(Ax, x)|, \quad x \in H.$$

Pour tout opérateur linéaire borné A se vérifient les inégalités: $\mu(A)\,\|A\| \leqslant \bar{\rho}(A) \leqslant \|A\|$, $\mu(A) \geqslant 1/2$, de plus, on a pour tout n naturel $\bar{\rho}(A^n) \leqslant [\bar{\rho}(A)]^n$. Si l'opérateur A est autoadjoint, on a $\bar{\rho}(A) = \|A\|$. Notons encore une série de propriétés intéressantes du rayon numérique. C'est ainsi, par exemple, que $\bar{\rho}(A^*) = \bar{\rho}(A)$, $\bar{\rho}(A^*A) = \|A\|^2$. En outre, $\rho(A) \leqslant \bar{\rho}(A)$, où $\rho(A)$ est le rayon spectral déjà introduit de l'opérateur.

L'opérateur linéaire A agissant dans l'espace hilbertien H est dit *positif* ($A > 0$) si $(Ax, x) > 0$ pour tous les $x \in H$, sauf pour $x = 0$. En cas d'espace complexe H, la définition de la positivité n'est introduite que pour les opérateurs autoadjoints, car la positivité de l'opérateur implique dans ce cas que ce dernier est aussi autoadjoint.

De façon analogue est introduite la définition de la *non-négativité* de l'opérateur A (pour tous $x \in H$ $(Ax, x) \geqslant 0$) et de la *définissabilité positive* (pour tous $x \in H$ $(Ax, x) \geqslant \delta(x, x)$, où $\delta > 0$).

L'opérateur non linéaire A agissant dans H est dit *monotone* si

$$(Ax - Ay,\ x - y) \geqslant 0,\quad x,\ y \in H,$$

rigoureusement monotone si

$$(Ax - Ay,\ x - y) > 0,\quad x,\ y \in H,\ x \neq y,$$

et *fortement monotone* si pour tous $x, y \in H$ on a l'inégalité

$$(Ax - Ay,\ x - y) \geqslant \delta\,\|x - y\|^2,\qquad \delta > 0.$$

Théorème 2. *Soit un opérateur non linéaire A possédant en chaque point $x \in H$ une dérivée Gateau continue. Dans ce cas l'opérateur A est fortement monotone sur H seulement et rien que seulement s'il existe un tel $\delta > 0$ pour lequel*

$$(A'(x)\,y,\ y) \geqslant \delta\,(y,\ y),\qquad y \in H.$$

Soit A un opérateur linéaire non négatif. Appelons le nombre (Ax, x) *énergie de l'opérateur*. Comparons en énergie les opérateurs A et B. Si $((A - B)\,x, x) \geqslant 0$ pour tous $x \in H$, on peut alors écrire $A \geqslant B$.

S'il existe des constantes telles que $\gamma_2 \geqslant \gamma_1 > 0$ entraînant pour les opérateurs linéaires A et B les inégalités $\gamma_1 B \leqslant A \leqslant \gamma_2 B$, on dira alors que ces opérateurs sont *énergétiquement équivalents* (én. éq.), γ_1 et γ_2 étant des constantes d'équivalence énergétique des opérateurs A et B. Posons

$$\delta = \inf_{\|x\|=1} (Ax,\ x)\ \text{ et }\ \Delta = \sup_{\|x\|=1} (Ax,\ x).$$

Les nombres δ et Δ sont appelés *bornes* de l'opérateur A (autoadjoint au cas de H complexe). Apparemment, les inégalités

$$\delta\,(x,\ x) \leqslant (Ax,\ x) \leqslant \Delta\,(x,\ x),\quad x \in H$$

ou

$$\delta E \leqslant A \leqslant \Delta E,$$

où E est un opérateur identique, $Ex = x$, se vérifient.

On se convainc sans peine que la relation d'inégalité introduite sur l'ensemble des opérateurs agissant dans H possède les propriétés suivantes:

1) de $A \geqslant B$ et $C \geqslant D$ il s'ensuit que $A + C \geqslant B + D$,
2) de $A \geqslant 0$ et $\lambda \geqslant 0$ il s'ensuit que $\lambda A \geqslant 0$,
3) de $A \geqslant B$ et $B \geqslant C$ il s'ensuit que $A \geqslant C$,
4) si $A > 0$ et A^{-1} existe, alors $A^{-1} > 0$.

Ensuite, il est évident que A^*A et AA^* sont des opérateurs non négatifs pour tout opérateur linéaire A. Ces opérateurs seront positifs si A est un opérateur positif.

Théorème 3. *Le produit AB des opérateurs A et B permutables non négatifs dont l'un est autoadjoint est également un opérateur non négatif.*

Pour un opérateur A autoadjoint non négatif quelconque a lieu l'inégalité généralisée de Cauchy-Bouniakovski

$$|(Ax, y)| \leqslant \sqrt{(Ax, x)}\sqrt{(Ay, y)}, \quad x, y \in H.$$

Soit D l'opérateur autoadjoint positif agissant dans H. On peut alors introduire l'*espace énergétique* H_D composé d'éléments de H avec produit scalaire $(x. y)_D = (Dx, y)$ et la norme

$$\|x\|_D = \sqrt{(Dx, x)}.$$

Notons que si D est un opérateur autoadjoint défini positif et borné dans H, alors, en vertu de l'inégalité de Cauchy-Bouniakovski, pour tout $x \in H$ se vérifient les estimations

$$\delta(x, x) \leqslant (Dx, x) \leqslant \|Dx\| \, \|x\| \leqslant \Delta(x, x), \ \Delta = \|D\|, \ \delta > 0.$$

Ces inégalités peuvent être transcrites en la forme

$$\sqrt{\delta}\|x\| \leqslant \|x\|_D \leqslant \sqrt{\Delta}\|x\|,$$

d'où il s'ensuit que la norme ordinaire $\|\cdot\|$ et la norme énergétique $\|\cdot\|_D$ sont équivalentes.

Remarquons que l'espace énergétique unitaire H_D peut être construit sur la base de l'opérateur positif non autoadjoint D. Pour cela le produit scalaire dans H_D sera défini de la façon suivante:

$$(x, y)_D = (D_0 x, y), \text{ où } D_0 = 0{,}5\,(D + D^*).$$

Fournissons une série de lemmes contenant les principales inégalités qui nous seront nécessaires dans la suite.

Lemme 6. *Supposons que pour l'opérateur linéaire est remplie la condition $A \geqslant \delta E$, $\delta > 0$. Dans ce cas pour tout $x \in H$ a lieu l'inégalité*

$$(Ax, Ax) \geqslant \delta(Ax, x).$$

Si pour un opérateur autoadjoint non négatif est remplie la condition $A \leqslant \Delta E$, alors pour tout $x \in H$ on a l'inégalité

$$(Ax, Ax) \leqslant \Delta (Ax, x).$$

L e m m e 7. *A partir de la condition $(Ax, Ax) \leqslant \Delta (Ax, x)$, $x \in H$, $\Delta > 0$, imposée à l'opérateur A non négatif, s'ensuit l'inégalité*

$$A \leqslant \Delta E,$$

tandis que de la condition $(Ax, Ax) \geqslant \delta (Ax, x)$, $\delta > 0$, imposée à l'opérateur autoadjoint non négatif A s'ensuit l'inégalité

$$A \geqslant \delta E.$$

C o r o l l a i r e 1. *Il s'ensuit des lemmes 6 et 7 que dans le cas de l'opérateur autoadjoint défini positif A les inégalités*

$$\delta E \leqslant A \leqslant \Delta E, \ \delta > 0$$

et

$$\delta (Ax, x) \leqslant (Ax, Ax) \leqslant \Delta (Ax, x), \quad \delta > 0,$$

sont équivalentes.

C o r o l l a i r e 2. *A partir de (5) et du lemme 6 s'ensuit l'estimation $(Ax, Ax) \leqslant \| A \| (Ax, x)$, $x \in H$ pour l'opérateur non négatif autoadjoint A dans H.*

L e m m e 8. *Soit A l'opérateur autoadjoint positif borné dans H tel que $A > 0$, $\| Ax \| \leqslant \Delta \| x \|$. Dans ce cas l'opérateur inverse A^{-1} est alors défini positif $A^{-1} \geqslant \frac{1}{\Delta} E$.*

L e m m e 9. *Soient A et B des opérateurs autoadjoints définis positifs dans H. Les inégalités*

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \quad \gamma_2 \geqslant \gamma_1 > 0$$

et

$$\gamma_1 A^{-1} \leqslant B^{-1} \leqslant \gamma_2 A^{-1}, \quad \gamma_2 \geqslant \gamma_1 > 0$$

sont alors équivalentes.

L e m m e 10. *Si A est un opérateur défini positif $A \geqslant \delta E$, $\delta > 0$, il existe alors un opérateur inverse A^{-1} et $\| A^{-1} \| \leqslant 1/\delta$.*

La démonstration s'ensuit de l'inégalité

$$\delta \| x \|^2 \leqslant (Ax, x) \leqslant \| Ax \| \, \| x \|, \ \delta > 0$$

et du théorème 1.

R e m a r q u e. Si A est un opérateur positif, A^{-1} existe alors. Au cas d'un espace H complexe, pour l'existence de A^{-1} il suffit que la composante réelle $A_0 = 0{,}5 (A + A^*)$ soit positive ou que soit positive la composante imaginaire $A_1 = \frac{1}{2i} (A - A^*)$ de l'opérateur A.

4. Fonctions de l'opérateur borné. En théorie des méthodes itératives on sera obligé d'aborder les fonctions de l'opérateur. Soit A l'opérateur linéaire borné agissant dans un espace normé X. Si $f(\lambda)$ est une fonction analytique entière de la variable λ se developpant en la série $\sum_{k=0}^{\infty} a_k\lambda^k$, il est alors possible de définir *la fonction $f(A)$ de l'opérateur A* au moyen de la formule $f(A) = \sum_{k=0}^{\infty} a_k A^k$. L'opérateur $f(A)$ sera également linéaire et borné. En guise d'exemple donnons la fonction exponentielle de l'opérateur $e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!}$. La définition introduite de la fonction de l'opérateur peut être généralisée à une classe plus étendue de fonctions et, ensuite, on peut bâtir le calcul opératoriel pour des opérateurs bornés. On ne donnera une définition plus généralisée que pour des opérateurs autoadjoints bornés dans l'espace hilbertien.

Soient δ et Δ les bornes inférieure et supérieure de l'opérateur A autoadjoint dans H. Soit $f(\lambda)$ une fonction continue sur le tronçon $[\delta, \Delta]$. L'opérateur $f(A)$ est appelé *fonction de l'opérateur autoadjoint A*.

La correspondance entre les fonctions de la variable réelle et les fonctions de l'opérateur se caractérise par les propriétés suivantes:

1) Si $f(\lambda) = \alpha f_1(\lambda) + \beta f_2(\lambda)$, $f(A) = \alpha f_1(A) + \beta f_2(A)$.

2) Si $f(\lambda) = f_1(\lambda) f_2(\lambda)$, $f(A) = f_1(A) f_2(A)$.

3) Il s'ensuit de $AB = BA$ que $f(A) B = Bf(A)$ pour tout opérateur B linéaire borné.

4) Si $f_1(\lambda) \leqslant f(\lambda) \leqslant f_2(\lambda)$ pour tous les $\lambda \in [\delta, \Delta]$, on a alors $f_1(A) \leqslant f(A) \leqslant f_2(A)$.

5) $\| f(A) \| \leqslant \max_{\delta \leqslant \lambda \leqslant \Delta} | f(\lambda) |$.

6) $\overline{f}(A) = [f(A)]^*$, où le trait au-dessus de la fonction indique le passage à la fonction complexe conjuguée. Si $f(\lambda)$ est une fonction réelle, il en suit que l'opérateur $f(A)$ est autoadjoint dans H.

Il s'ensuit de la propriété 4) que si $f(\lambda) \geqslant 0$ sur $[\delta, \Delta]$, $f(A)$ est un opérateur non négatif.

Un exemple important de fonction de l'opérateur est la racine carrée de l'opérateur. L'opérateur B est appelé *racine carrée de l'opérateur A* si $B^2 = A$.

Théorème 4. *Il existe une racine carrée autoconjuguée non négative unique de l'opérateur autoadjoint quelconque non négatif A permutable avec tout opérateur permutable avec A.*

La racine carrée de l'opérateur A sera désignée par $A^{1/2}$. Notons la propriété suivante: $\| A \| = \| A^{1/2} \|^2$ si $A = A^* \geqslant 0$.

Théorème 5. *Si A est un opérateur autoadjoint défini positif, $A = A^* \geqslant \delta E$, $\delta > 0$, il existe alors un opérateur autoadjoint borné $A^{-1/2}$, $\| A^{-1/2} \| \leqslant 1/\sqrt{\delta}$.*

La démonstration s'ensuit de l'inégalité

$$\delta (x, x) \leqslant (Ax, x) = (A^{1/2}x, A^{1/2}x) = \| A^{1/2}x \|^2$$

et du théorème 1.

5. Opérateurs dans un espace de dimension finie. Examinons l'espace unitaire H à n dimensions. Soient $x_1, x_2, \ldots, x_n$ les éléments composant la base orthonormée dans H. Selon la définition de l'espace de dimension finie tout élément $x \in H$ peut être représenté de façon unique sous forme d'une combinaison linéaire

$$x = c_1x_1 + c_2x_2 + \ldots + c_nx_n. \quad (6)$$

Il s'ensuit de l'orthonormalité du système $x_1, x_2, \ldots, x_n$ que $c_k = (x, x_k)$.

Donc à chaque élément $x \in H$ on peut faire correspondre le vecteur $c = (c_1, c_2, \ldots, c_n)$ dont les composantes sont les coefficients c_k du développement (6).

Soit A l'opérateur linéaire donné sur H. Il lui correspond dans la base $x_1, x_2, \ldots, x_n$ la matrice $\mathcal{A} = (a_{ik})$ de dimension $n \times n$, où $a_{ik} = (Ax_k, x_i)$. Inversement, toute matrice $\mathcal{A}$ de dimension $n \times n$ définit l'opérateur linéaire dans H. Dans ce cas à l'élément Ax on fait correspondre le vecteur $(\sum_{k=1}^{n} a_{1k}c_k, \sum_{k=1}^{n} a_{2k}c_k, \ldots$ $\ldots, \sum_{k=1}^{n} a_{nk}c_k)$, c'est-à-dire le vecteur $\mathcal{A}c$.

Si l'opérateur A est autoadjoint dans H, la matrice $\mathcal{A}$ qui lui est associée est symétrique dans toute base orthonormée. Notons que dans une base non orthonormée à l'opérateur autoadjoint A correspond une matrice asymétrique.

Arrêtons-nous sur les propriétés des valeurs propres et des éléments propres de l'opérateur linéaire A. Le nombre λ est appelé *valeur propre de l'opérateur A* si l'équation

$$Ax = \lambda x \quad (7)$$

possède des solutions non nulles. L'élément $x \neq 0$ satisfaisant à (7) est appelé *élément propre de l'opérateur A* associé à la valeur propre λ. Autrement dit, les valeurs propres de l'opérateur A ce sont les valeurs de λ pour lesquelles $\ker (A - \lambda E) \neq 0$; les éléments propres correspondant à la valeur propre λ sont des éléments différant de zéro du sous-espace $\ker (A - \lambda E)$. Quant à ce sous-espace, il est appelé *sous-espace propre* associé à la valeur propre λ.

L'ensemble $\sigma (A)$ de valeurs propres de l'opérateur A est dénommé *spectre de l'opérateur A*.

1. L'opérateur autoadjoint A possède n éléments propres orthonormés $x_1, x_2, \ldots, x_n$. Les valeurs propres associées λ_k, $k = 1, 2, \ldots, n$ sont réelles; Si toutes les valeurs propres sont différentes, A est dit *opérateur à simple spectre*.

2. Pour un opérateur autoadjoint A il y a lieu aux égalités

$$\| A \| = \rho(A) = \max_{1 \leqslant k \leqslant n} | \lambda_k |,$$

où $\rho(A)$ est le *rayon spectral de l'opérateur* A. Ces égalités se conservent également pour un opérateur normal A.

3. Si $A = A^* \geqslant 0$, toutes les valeurs propres de l'opérateur A sont non négatives. Dans ce cas pour tout $x \in H$

$$\delta(x, x) \leqslant (Ax, x) \leqslant \Delta(x, x),$$

où $0 \leqslant \delta = \min\limits_k \lambda_k$, $\Delta = \max\limits_k \lambda_k$. On appelle *quotient de Rayleigh* l'expression $(Ax, x)/(x, x)$ associée à l'opérateur autoadjoint.

Les valeurs propres minimale et maximale de l'opérateur A se déterminent à l'aide du quotient de Rayleigh de la façon suivante:

$$\delta = \min_{x \neq 0} \frac{(Ax, x)}{(x, x)}, \quad \Delta = \max_{x \neq 0} \frac{(Ax, x)}{(x, x)}.$$

4. Notons par $\lambda(A)$ les valeurs propres de l'opérateur A. Soit $f(A)$ la fonction de l'opérateur autoadjoint A. On a alors $\lambda(f(A)) = f(\lambda(A))$ (théorème de l'application des spectres).

5. Si les opérateurs autoadjoints A et B sont permutables, $A = A^*$, $B = B^*$, $AB = BA$, ils possèdent alors un système commun d'éléments propres. De plus, les opérateurs AB et $A + B$ ont le même système d'éléments propres que les opérateurs A et B, et les valeurs propres

$$\lambda(AB) = \lambda(A)\lambda(B), \quad \lambda(A + B) = \lambda(A) + \lambda(B).$$

6. Un élément quelconque $x \in H$ peut être développé en éléments propres de l'opérateur autoadjoint A

$$x = \sum_{k=1}^{n} c_k x_k, \; c_k = (x, x_k). \quad \text{avec} \quad \| x \|^2 = \sum_{k=1}^{n} c_k^2.$$

Le nombre λ est appelé *valeur propre de l'opérateur* A *relativement à l'opérateur* B si l'équation

$$Ax = \lambda Bx \tag{8}$$

possède des solutions non nulles. L'élément $x \neq 0$ vérifiant l'équation (8) est appelé *élément propre de l'opérateur* A *relativement à l'opérateur* B, associé au nombre λ.

7. Si les opérateurs A et B sont autoadjoints dans H, tandis que l'opérateur B est, de plus, défini positif, il existe n éléments propres $x_1, x_2, \ldots, x_n$ orthonormés dans l'espace énergétique H_B: $(x_k, x_i)_B = \delta_{ki}$, $k, i = 1, 2, \ldots, n$. Les valeurs propres associées sont

réelles et on a les inégalités

$$\gamma_1 (Bx, x) \leqslant (Ax, x) \leqslant \gamma_2 (Bx, x),$$

où

$$\gamma_1 = \min_k \lambda_k = \min_{x \neq 0} \frac{(Ax, x)}{(Bx, x)},$$

$$\gamma_2 = \max_k \lambda_k = \max_{x \neq 0} \frac{(Ax, x)}{(Bx, x)}.$$

Donc les constantes des opérateurs autoadjoints én. éq. A et B au cas où l'opérateur B est défini positif coïncident avec les valeurs propres minimale et maximale du problème généralisé (8).

6. Résolubilité des équations opératorielles. Supposons qu'il s'agit de trouver la solution de l'équation opératorielle de première espèce

$$Au = f, \tag{9}$$

où A est un opérateur linéaire borné dans l'espace hilbertien H, f l'élément donné et u l'élément cherché de H. Posons que H est de dimension finie. On s'intéressera au problème de la résolubilité de l'équation (9). On a le théorème:

Théorème 6. *Pour que l'équation* (9) *soit résoluble pour tout second membre* f, *il faut et il suffit que l'équation homogène correspondante* $Au = 0$ *ne possède qu'une solution triviale* $u = 0$. *De plus, la solution de l'équation* (9) *est unique.*

La démonstration du théorème se fonde sur le lemme 1.

On peut formuler le théorème d'une autre manière: l'équation (9) se résout d'une façon unique pour tout $f \in H$ seulement et rien que seulement quand $\ker A = 0$ (voir point 2).

Si $\ker A \neq 0$, l'équation ne se résout qu'avec une limitation supplémentaire sur f. Rappelons qu'en vertu du lemme 3 l'espace H est une somme directe des sous-espaces orthogonaux: $H = \ker A \oplus \oplus \operatorname{im} A^*$, $H = \ker A^* \oplus \operatorname{im} A$.

Théorème 7. *Pour la résolubilité de l'équation inhomogène* (9) *il faut et il suffit que le second membre* f *soit orthogonal au sous-espace* $\ker A^*$. *Dans ce cas la solution n'est pas unique et est déterminée à la précision de l'élément arbitraire près appartenant à* $\ker A$:

$$u = \tilde{u} + \bar{u}, \quad \tilde{u} \in \ker A, \quad A\bar{u} = f, \quad \bar{u} \in \operatorname{im} A^*.$$

Posons f orthogonal à $\ker A^*$. On appelle *solution normale de* (9) la solution ayant une norme minimale.

Lemme 11. *Une solution normale est unique et appartient au sous-espace* $\operatorname{im} A^*$ (*c'est-à-dire est orthogonale à* $\ker A$).

En effet, soit $u = \tilde{u} + \bar{u}$, $\tilde{u} \in \ker A$, $\bar{u} \in \operatorname{im} A^*$. Dans ce cas

$\| u \|^2 = (u, u) = \| \tilde{u} \|^2 + \| \bar{u} \|^2$, vu que $\tilde{u}$ est un élément quelconque du sous-espace $\ker A$. Donc la norme $\| u \|$ sera minimale si $u = \bar{u} \in \operatorname{im} A^*$.

Supposons que la condition d'orthogonalité de f au sous-espace $\ker A^*$ n'est pas remplie. Dans ce cas la solution de l'équation (9) au sens classique n'existe pas. Soit

$$f = \tilde{f} + \bar{f}, \quad \tilde{f} \in \ker A^*, \quad \bar{f} \in \operatorname{im} A.$$

Par *solution généralisée de l'équation* (9) on entend l'élément $u \in H$ pour lequel $Au = \bar{f}$; la solution généralisée garantit un minimum à la fonctionnelle $\| Au - f \|$. En effet, puisque $(Au - \bar{f}) \in \operatorname{im} A$ pour tout $u \in H$, on a

$$\| Au - f \|^2 = \| Au - \bar{f} \|^2 + \| \tilde{f} \|^2 \geqslant \| \tilde{f} \|^2,$$

l'égalité se vérifiant si u est une solution généralisée.

La solution généralisée est définie à la précision de l'élément quelconque du sous-espace $\ker A$ près. Appelons solution généralisée normale de l'équation (9) la solution généralisée présentant une norme minimale. La solution normale est unique et appartient à $\operatorname{im} A^*$.

La notion de solution normale introduite ici est apparemment en accord avec celle fournie plus haut. Notons que si l'on est en présence d'une solution normale classique, cette dernière coïncide avec la solution normale généralisée.

Examinons maintenant l'équation (9) munie de l'opérateur non linéaire arbitraire A agissant dans l'espace hilbertien H. Dans ce cas, pour démontrer l'existence et l'unicité de la solution de l'équation (9), on recourt souvent *au principe des applications contractantes de S. Banach.*

Théorème 8. *Soit donné dans un espace hilbertien H l'opérateur B, application de l'ensemble fermé T de l'espace H en lui-même. Supposons en outre que l'opérateur B est de contraction régulière, c'est-à-dire satisfaisant à la condition de Lipschitz*

$$\| Bx - By \| \leqslant q \, \| x - y \|, \quad x, y \in T,$$

où $q < 1$ et ne dépend pas de x et y. Il existe alors un point et un seul $x_* \in T$ pour lequel $x_* = Bx_*$.

Le point x_* est dit *point immobile de l'opérateur B.*

Corollaire 1. *Si l'opérateur B possède une dérivée Gateau dans H qui vérifie la condition $\| B'(x) \| \leqslant q < 1$ pour tout $x \in H$, l'équation $x = Bx$ possède en H une solution unique.*

Corollaire 2. *Soit C l'opérateur constituant une application de l'ensemble fermé T en lui-même et qui commute avec l'opérateur B satisfaisant aux conditions du principe des applications contractantes. Alors le point immobile de l'opérateur B est un point immobile (vrai-*

semblablement non unique) de l'opérateur C. En particulier, si par une certaine itération B^n de l'opérateur B on satisfait au principe des applications contractantes, le point immobile de l'opérateur B^n est également un point immobile (unique) de l'opérateur B.

Revenons maintenant à la solution de l'équation (9) munie de l'opérateur non linéaire A. Il y a lieu au

Théorème 9. *Admettons que l'opérateur A possède en chaque point $x \in H$ une dérivée Gateau $A'(x)$ et qu'il existe un $\tau \neq 0$ pour lequel pour tous les $x \in H$ est vérifiée l'estimation $\| E - \tau A'(x) \| \leqslant \leqslant q < 1$. L'équation (9) possède dans ce cas dans H une solution unique.*

En effet, l'équation (9) peut être écrite sous la forme suivante:

$$u = u - \tau A u + \tau f, \quad \tau \neq 0. \tag{10}$$

Définissons l'opérateur B: $Bx = x - \tau Ax + \tau f$. L'opérateur B a apparemment une dérivée Gateau égale à $B'(x) = E - \tau A'(x)$. En vertu des conditions du théorème, on a $\| B'(x) \| \leqslant q < 1$ pour tout $x \in H$. Aussi à partir du corollaire 1 du théorème 8 déduit-on l'existence et l'unicité de la solution de l'équation (10) et, partant, de l'équation (9). Le théorème est démontré.

Remarquons que dans le ch. VI seront étudiés certains procédés d'obtention des estimations de normes pour les opérateurs linéaires de la forme $E - \tau C$, où τ est un nombre.

Le principe des applications contractantes ne couvre pas tous les cas d'existence de la solution d'une équation non linéaire. Dans la démonstration de la résolubilité de l'équation opératorielle (9) on peut utiliser l'une des variantes du théorème sur le point immobile, le *principe de Brouyder.*

Théorème 10. *Soit dans un espace hilbertien de dimension finie H un opérateur continu et monotone (rigoureusement monotone) B qui vérifie la condition*

$$(Bx, x) \geqslant 0 \text{ pour } \| x \| = \rho > 0.$$

Alors l'équation $Bx = 0$ possède dans une sphère $\| x \| \leqslant \rho$ au moins une (et, partant, unique) solution.

Utilisons ce théorème et formulons les conditions avec la satisfaction desquelles l'équation opératorielle (9) est résoluble de façon unique pour tout second membre f.

Théorème 11. *Soit donnée dans un espace hilbertien de dimension finie H l'équation (9) munie d'un opérateur continu A fortement monotone*

$$(Ax - Ay, x - y) \geqslant \delta \| x - y \|^2, \quad \delta > 0, \ x, y \in H.$$

Dans ce cas l'équation (9) possède dans la sphère $\| u \| \leqslant \frac{1}{\delta} \| A0 - f \|$ une solution unique.

En effet, écrivons l'équation (4) sous la forme suivante:

$$Bu = Au - f = 0.$$

On constate que l'opérateur B est continu et fortement monotone. En utilisant la condition du théorème et l'inégalité de Cauchy-Bouniakovski, il vient

$$(Bx, x) = (Ax - f, x) = (Ax - A0, x - 0) - (f - A0, x) \geqslant$$
$$\geqslant \delta \|x\|^2 - \|f - A0\| \|x\| = (\delta \|x\| - \|A0 - f\|) \|x\|.$$

Il s'ensuit que sur la sphère $\|x\| = \frac{1}{\delta} \|A0 - f\|$ l'opérateur B satisfait à la condition $(Bx, x) \geqslant 0$. Aussi en vertu du théorème 10 l'équation $Bu = 0$ (et avec elle l'équation (9)) admet-elle une solution unique dans la sphère considérée. Le théorème 11 est démontré.

Corollaire 1. *Si l'opérateur A possède dans H une dérivée Gateau qui est un opérateur défini positif dans H, alors les conditions du théorème* 11 *sont satisfaites.*

En effet, comme dans l'espace de dimension finie l'opérateur est borné, la dérivée Gateau y est un opérateur continu borné et défini positif dans H. Il s'ensuit du théorème 2 que A est un opérateur fortement monotone. En outre, du fait que la dérivée Gateau est bornée, il s'ensuit que l'opérateur A satisfait à la condition de Lipschitz et est donc continu.

§ 2. Schémas aux différences considérés comme des équations opératorielles

1. Exemples d'espaces de fonctions de mailles. On a introduit au § 1, ch. 1 les principales notions de la théorie des schémas aux différences finies: maillages, équations de mailles, fonctions de mailles, différences divisées, etc. La théorie formule les principes généraux et les règles de mise en œuvre de schémas aux différences de qualité établie. Le trait caractéristique de cette théorie est la possibilité d'opposer à chaque équation différentielle une classe entière de schémas aux différences jouissant des propriétés exigées. Il est naturel de vouloir se libérer de la structure concrète et de la forme explicite des équations aux différences lors de la construction de la théorie générale. On est ainsi amené à définir les schémas aux différences comme des équations opératorielles aux opérateurs agissant dans un certain espace fonctionnel, à savoir l'espace de fonctions de mailles.

Par l'*espace de fonctions de mailles* on entend un ensemble de fonctions données sur un certain maillage. Comme à chaque fonction de maille on peut faire correspondre un vecteur dont les coordonnées sont des valeurs de la fonction de maille aux nœuds du maillage, les opérations d'addition des fonctions et de multiplication des

fonctions par un nombre se définissent de la même manière que pour les vecteurs.

L'espace des fonctions de mailles est linéaire et si le maillage est composé d'un nombre fini de nœuds l'espace est de dimension finie. Cette dimension est égale au nombre de nœuds du maillage.

Dans l'espace de fonctions de mailles on peut introduire le produit scalaire des fonctions en rendant cet espace hilbertien. Les espaces variés de fonctions de mailles diffèrent l'un de l'autre par le choix du maillage et le type de normalisation. Donnons quelques exemples.

E x e m p l e 1. Soit sur le segment $0 \leqslant x \leqslant l$ un maillage régulier $\bar{\omega} = \{x_i = ih,\ 0 \leqslant i \leqslant N,\ hN = l\}$ de pas h. Désignons par ω, ω^+ et ω^- les parties suivantes du maillage $\bar{\omega}$:

$$\omega = \{x_i \in \bar{\omega},\ 1 \leqslant i \leqslant N-1\},$$

$$\omega^+ = \{x_i \in \bar{\omega},\ 1 \leqslant i \leqslant N\},$$

$$\omega^- = \{x_i \in \bar{\omega},\ 0 \leqslant i \leqslant N-1\}.$$

Sur l'ensemble H des fonctions de mailles données sur $\bar{\omega}$ et prenant des valeurs réelles déterminons le produit scalaire et la norme de la façon suivante:

$$(u, v) = (u, v)_{\bar{\omega}} = \sum_{i=1}^{N-1} u_i v_i h + 0{,}5h\,(u_0 v_0 + u_N v_N),$$
$$\|u\| = \sqrt{(u, u)},\quad u_i = u(x_i),\quad v_i = v(x_i). \tag{1}$$

Si l'on considère u_i et v_i comme des valeurs sur le maillage $\bar{\omega}$ des fonctions $u(x)$ et $v(x)$ de l'argument continu $x \in [0, l]$, le produit scalaire (1) constitue alors la formule de quadrature des trapèzes de l'intégrale $\int_0^l u(x)\, v(x)\, dx$. Si les fonctions de mailles sont données sur ω, ω^+ ou ω^-, le produit scalaire des fonctions de mailles réelles s'obtient respectivement suivant les formules

$$(u, v) = \sum_{i=1}^{N-1} u_i v_i h,\quad u, v \in H(\omega),$$

$$(u, v) = \sum_{i=1}^{N-1} u_i v_i h + 0{,}5h u_N v_N,\quad u, v \in H(\omega^+),$$

$$(u, v) = \sum_{i=1}^{N-1} u_i v_i h + 0{,}5h u_0 v_0,\quad u, v \in H(\omega^-).$$

On vérifie sans peine que les produits scalaires introduits satisfont à tous les axiomes du produit scalaire et, par suite, les espaces construits sont hilbertiens.

Exemple 2. Introduisons maintenant sur le tronçon $0 \leqslant x \leqslant l$ un maillage irrégulier quelconque

$$\bar{\omega} = \{x_i \in [0, l], \; x_i = x_{i-1} + h_i, \; 1 \leqslant i \leqslant N, \; x_0 = 0, \; x_N = l\}. \quad (2)$$

Rappelons la définition du pas moyen $\hbar_i$ au nœud x_i:

$$\hbar_i = 0{,}5\,(h_i + h_{i+1}), \; 1 \leqslant i \leqslant N-1, \; \hbar_0 = 0{,}5h_1, \; \hbar_N = 0{,}5h_N. \quad (3)$$

Notons qu'un maillage régulier est un cas particulier du maillage irrégulier (2) pour $h_i \equiv h$. On a dans ce cas $\hbar_i = h$, $1 \leqslant i \leqslant N-1$, $\hbar_0 = \hbar_N = 0{,}5h$.

Désignons, comme plus haut, au moyen de ω, ω^+ et ω^- les parties respectives du maillage $\bar{\omega}$. Par analogie avec l'exemple 1, définissons dans les espaces réels des fonctions de mailles données sur les maillages considérés le produit scalaire suivant les formules:

$$(u, v) = \sum_{i=0}^{N} u_i v_i \hbar_i, \quad u, v \in H(\bar{\omega}), \quad (4)$$

$$(u, v) = \sum_{i=1}^{N-1} u_i v_i \hbar_i, \quad u, v \in H(\omega), \quad (5)$$

$$(u, v) = \sum_{i=1}^{N} u_i v_i \hbar_i, \quad u, v \in H(\omega^+),$$

$$(u, v) = \sum_{i=0}^{N-1} u_i v_i \hbar_i, \quad u, v \in H(\omega^-).$$

Les espaces des fonctions de mailles ainsi construits sont *hilbertiens* et possèdent une *dimension finie* égale au nombre de nœuds du maillage correspondant.

Les produits scalaires introduits peuvent être écrits sous la forme

$$(u, v) = \sum_{x_i \in \Omega} u(x_i)\, v(x_i)\, \hbar(x_i), \quad u, v \in H(\Omega),$$

où par Ω on entend soit $\bar{\omega}$, soit ω, ω^+ ou ω^-. Outre les produits scalaires mentionnés, on rencontre des sommes sous forme de

$$(u, v)_{\omega^+} = \sum_{i=1}^{N} u_i v_i h_i, \quad (u, v)_{\omega^-} = \sum_{i=0}^{N-1} u_i v_i h_{i+1}, \quad (6)$$

qui peuvent être utilisées en guise de produits scalaires dans les espaces $H(\omega^+)$ et $H(\omega^-)$. On constate que pour le produit scalaire (4) dans l'espace $H(\bar{\omega})$ se vérifie l'égalité

$$(u, v) = 0{,}5\,[(u, v)_{\omega^+} + (u, v)_{\omega^-}], \quad u, v \in H(\bar{\omega}).$$

Exemple 3. Soit un maillage rectangulaire irrégulier quelconque $\bar{\omega} = \bar{\omega}_1 \times \bar{\omega}_2$, où

$$\bar{\omega}_\alpha = \{x_\alpha(i_\alpha) \in [0, l_\alpha], \ x_\alpha(i_\alpha) = x_\alpha(i_\alpha - 1) + h_\alpha(i_\alpha),$$
$$1 \leqslant i_\alpha \leqslant N_\alpha, \ x_\alpha(0) = 0, \ x_\alpha(N_\alpha) = l_\alpha\}, \ \alpha = 1, 2,$$

introduit dans le rectangle $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha, \ \alpha = 1, 2\}$. Soit $\hbar_\alpha(i_\alpha)$, $0 \leqslant i_\alpha \leqslant N_\alpha$ le pas moyen au nœud $x_\alpha(i_\alpha)$ en direction de x_α:

$$\hbar_\alpha(i_\alpha) = 0{,}5\,[h_\alpha(i_\alpha) + h_\alpha(i_\alpha + 1)], \qquad 1 \leqslant i_\alpha \leqslant N_\alpha - 1,$$
$$\hbar_\alpha(0) = 0{,}5h_\alpha(1), \ \hbar_\alpha(N_\alpha) = 0{,}5h_\alpha(N_\alpha), \ \alpha = 1, 2.$$

Dans l'espace $H(\Omega)$ des fonctions de mailles données sur Ω, où Ω est une partie quelconque du maillage $\bar{\omega}$, déterminons le produit scalaire suivant la formule

$$(u, v) = \sum_{x_i \in \Omega} u(x_i)\, v(x_i)\, \hbar_1 \hbar_2, \qquad x_i = (x_1(i_1), \quad x_2(i_2)).$$

En particulier, si le maillage est régulier dans chaque direction, $h_\alpha(i_\alpha) \equiv h_\alpha$, $\alpha = 1, 2$, et les fonctions de mailles sont données sur ω (aux nœuds intérieurs du maillage $\bar{\omega}$), le produit scalaire introduit s'écrit sous la forme

$$(u, v) = \sum_{i_1=1}^{N_1-1} \sum_{i_2=1}^{N_2-1} u(i_1, i_2)\, v(i_1, i_2)\, h_1 h_2, \qquad u, v \in H(\omega).$$

On se limitera ici aux exemples donnés, quant aux autres exemples, plus compliqués, ils seront étudiés dans les chapitres suivants avec l'analyse de problèmes de différences concrets.

2. Quelques identités au sens de différences finies. Passons maintenant à la déduction des formules principales permettant de transformer les expressions renfermant les fonctions de mailles. On donnera ces formules pour le cas où les fonctions de mailles sont associées au maillage irrégulier défini dans (2).

Rappelons la définition des principales différences divisées de la fonction de maille:

$$y_{\bar{x}, i} = \frac{y_i - y_{i-1}}{h_i}, \qquad y_{x, i} = y_{\bar{x}, i+1} = \frac{y_{i+1} - y_i}{h_{i+1}}, \qquad y_{\check{x}, i} = \frac{y_i - y_{i-1}}{\hbar_i},$$
$$y_{\hat{x}, i} = \frac{y_{i+1} - y_i}{\hbar_i}, \qquad y_{\bar{x}\hat{x}, i} = y_{x\check{x}, i} = \frac{1}{\hbar_i}(y_{x, i} - y_{\bar{x}, i}).$$

Au point 2, § 1, ch. I on a obtenu deux formules de sommation par parties:

$$\sum_{i=m+1}^{n-1} u_{\hat{x},i} v_i \hbar_i = -\sum_{i=m+1}^{n} u_i v_{\bar{x},i} h_i + u_n v_n - u_{m+1} v_m, \tag{7}$$

$$\sum_{i=m+1}^{n-1} u_{\bar{x},i} v_i h_i = -\sum_{i=m}^{n-1} u_i v_{\hat{x},i} \hbar_i + u_{n-1} v_n - u_m v_m. \tag{8}$$

En portant dans ces formules les relations

$$h_i u_{\bar{x},i} = \hbar_i u_{\check{x},i}, \quad \hbar_i u_{\hat{x},i} = h_{i+1} u_{x,i},$$

après des transformations simples, on obtient les formules

$$\sum_{i=m+1}^{n-1} u_{\check{x},i} v_i \hbar_i = -\sum_{i=m}^{n-1} u_i v_{x,i} h_{i+1} + u_{n-1} v_n - u_m v_m, \tag{9}$$

$$\sum_{i=m+1}^{n-1} u_{x,i} v_i h_{i+1} = -\sum_{i=m+1}^{n} u_i v_{\check{x},i} \hbar_i + u_n v_n - u_{m+1} v_m, \tag{10}$$

$$\sum_{i=m+1}^{n} u_{\bar{x},i} v_i h_i = -\sum_{i=m}^{n-1} u_i v_{x,i} h_{i+1} + u_n v_n - u_m v_m. \tag{11}$$

Portons dans les formules (7), (9), (11) $m = 0$ et $n = N$ en tenant compte de la définition (5) du produit scalaire dans $H(\omega)$ ainsi que des notations (6). On obtient les identités

$$(u_x, v) = -(u, v_{\bar{x}})_{\omega^+} + u_N v_N - u_1 v_0, \tag{7'}$$

$$(u_{\check{x}}, v) = -(u, v_x)_{\omega^-} + u_{N-1} v_N - u_0 v_0, \tag{9'}$$

$$(u_{\bar{x}}, v)_{\omega^+} = -(u, v_x)_{\omega^-} + u_N v_N - u_0 v_0 \tag{11'}$$

pour les fonctions de mailles u_i et v_i données sur le maillage $\bar{\omega}$. Si l'on pose dans (7′) $u_i = a_i y_{\bar{x},i}$ pour $1 \leqslant i \leqslant N$, on obtient alors la formule de Green au sens de différences finies

$$((a y_{\bar{x}})_{\check{x}}, v) = -(a y_{\bar{x}}, v_{\bar{x}})_{\omega^+} + a_N y_{\bar{x},N} v_N - a_1 y_{x,0} v_0. \tag{12}$$

De façon analogue, en posant dans (9) $u_i = a_i y_{x,i}$ pour $0 \leqslant i \leqslant N-1$, il vient

$$((a y_x)_{\check{x}}, v) = -(a y_x, v_x)_{\omega^-} + a_{N-1} y_{\bar{x},N} v_N - a_0 y_{x,0} v_0.$$

Si de (12) on ôte l'égalité

$$((y, (a v_{\bar{x}})_{\hat{x}}) = -(a y_{\bar{x}}, v_{\bar{x}})_{\omega^+} + a_N v_{\bar{x},N} y_N - a_1 x_{x,0} y_0,$$

on obtient la seconde formule de Green au sens de différences finies

$$((a y_{\bar{x}})_{\hat{x}}, v) - (y, (a v_{\bar{x}})_{\hat{x}}) = a_N (y_{\bar{x}} v - v_{\bar{x}} y)_N - a_1 (y_x v - v_x y)_0. \tag{13}$$

Notons que pour les fonctions y_i et v_i devenant nulles pour $i = 0$ et $i = N$ ($y_0 = y_N = 0$, $v_0 = v_N = 0$), la formule (12) prend la forme

$$((ay_{\bar{x}})_{\hat{x}}, v) = -(ay_{\bar{x}}, v_{\bar{x}})_{\omega^+},$$

tandis que la seconde formule de Green (13) devient

$$((ay_{\bar{x}})_{\hat{x}}, v) = (y, (av_{\bar{x}})_{\hat{x}}).$$

Dans le cas général de fonctions de mailles quelconques données sur $\bar{\omega}$, les formules (12) et (13) peuvent être écrites sous la forme

$$(\Lambda y, v) = -(ay_{\bar{x}}, v_{\bar{x}})_{\omega^+}, \quad (\Lambda y, v) - (y, \Lambda v) = 0, \tag{14}$$

où l'opérateur de différences Λ, application de $H(\bar{\omega})$ sur $H(\bar{\omega})$, se définit de la façon suivante:

$$\Lambda y_i = \begin{cases} \dfrac{1}{\hbar_0} a_1 y_{x,0}, & i = 0, \\ (ay_{\bar{x}})_{\hat{x},i}, & 1 \leqslant i \leqslant N-1, \\ -\dfrac{1}{\hbar_N} a_N y_{\bar{x},N}, & i = N. \end{cases}$$

Le produit scalaire dans $H(\bar{\omega})$ est ici donné par la formule (4). Notons que l'égalité (14) exprime que l'opérateur Λ est autoadjoint dans l'espace $H(\bar{\omega})$.

On a examiné le cas quand les fonctions de mailles prennent sur le maillage des valeurs réelles. Si ces dernières prennent sur $\bar{\omega}$ des valeurs complexes, il faut introduire l'espace hilbertien complexe $H(\bar{\omega})$ muni du produit scalaire

$$(u, v) = \sum_{i=0}^{N} u_i \bar{v}_i \hbar_i, \quad u, v \in H(\bar{\omega}), \tag{15}$$

où $\bar{v}_i$ est le nombre conjugué complexe de v_i. De façon analogue se détermine le produit scalaire dans $H(\omega)$

$$(u, v) = \sum_{i=1}^{N-1} u_i \bar{v}_i h_i, \quad u, v \in H(\omega), \tag{16}$$

de même que dans $H(\omega^+)$ et $H(\omega^-)$. De plus, les formules de sommation en parties (7′), (9′), (11′) prennent la forme

$$(u_{\hat{x}}, v) = -(u, v_{\bar{x}})_{\omega^+} + u_N \bar{v}_N - u_1 \bar{v}_0,$$

$$(u_{\check{x}}, v) = -(u, v_x)_{\omega^-} + u_{N-1} \bar{v}_N - u_0 \bar{v}_0,$$

$$(u_{\bar{x}}, v)_{\omega^+} = -(u, v_x)_{\omega^-} + u_N \bar{v}_N - u_0 \bar{v}_0.$$

tandis que les formules de Green au sens des différences finies la forme

$$((ay_{\bar{x}})_{\hat{x}},\ v) = -(ay_{\bar{x}},\ v_{\bar{x}})_{\omega^+} + a_N y_{\bar{x},\,N} \bar{v}_N - a_1 y_{x,\,0} \bar{v}_0,$$

$$((ay_{\bar{x}})_{\hat{x}},\ v) - (y,\ (av_{\bar{x}})_{\hat{x}}) = ((\bar{a} - a)\, y_{\bar{x}},\ v_{\bar{x}})_{\omega^+} +$$
$$+ (ay_{\bar{x}}\bar{v} - \bar{a}y\bar{v}_{\bar{x}})_N - (a_1 y_{x,\,0}\bar{v}_0 - \bar{a}_1 y_0 \bar{v}_{x,\,0}).$$

On a utilisé ici la notation (16).

En profitant de l'opérateur Λ introduit plus haut et de la notation (15) du produit scalaire dans $H(\bar{\omega})$, on est en mesure d'écrire la seconde formule de différences de Green sous la forme

$$(\Lambda y,\ v) - (y,\ \Lambda v) = ((\bar{a} - a)\, y_{\bar{x}},\ v_{\bar{x}})_{\omega^+}.$$

Il en suit que dans l'espace hilbertien complexe $H(\bar{\omega})$ l'opérateur Λ est autoadjoint si tous les a_i sont réels.

Les relations analogues aux première et seconde formules de différences de Green (12), (13) ont également lieu au cas de l'opérateur de différences $(ay_{\bar{x}\hat{x}})_{\bar{x}\hat{x}}$. Donnons, par exemple, l'analogue de la formule (12)

$$\sum_{i=2}^{N-2} (ay_{\bar{x}\hat{x}})_{\bar{x}\hat{x},\,i} v_i \hbar_i = \sum_{i=1}^{N-1} a_i y_{\bar{x}\hat{x},\,i} v_{\bar{x}\hat{x},\,i} \hbar_i +$$
$$+ [(ay_{\bar{x}\hat{x}})_{\bar{x}} v - ay_{\bar{x}\hat{x}} v_x]_{N-1} - [(ay_{\bar{x}\hat{x}})_x v - ay_{\bar{x}\hat{x}} v_{\bar{x}}]_1.$$

3. Bornes des opérateurs de différences les plus simples. Avec l'étude des propriétés des opérateurs de différences on s'est servi d'inégalités permettant d'apprécier les bornes des opérateurs et les constantes d'équivalence énergétique de deux opérateurs agissant dans l'espace de fonctions de mailles H.

Voyons d'abord les opérateurs de différences donnés sur un ensemble de fonctions de mailles d'un argument, définis sur un maillage régulier $\bar{\omega} = \{x_i = ih \in [0,\ l],\ 0 \leqslant i \leqslant N,\ hN = l\}$. On utilisera plus loin les notations

$$(u,\ v) = \sum_{i=1}^{N-1} u_i v_i h + 0{,}5h\,(u_0 v_0 + u_N v_N),\quad (u,\ v)_{\omega^+} = \sum_{i=1}^{N} u_i v_i h.$$

Il y a lieu au

L e m m e 12. *Pour toute fonction $y_i = y(x_i)$ donnée sur un maillage régulier $\bar{\omega}$ et devenant nulle pour $i = 0$ et $i = N$ se vérifient les inégalités*

$$\gamma_1 (y,\ y) \leqslant (y_{\bar{x}}^2,\ 1)_{\omega^+} \leqslant \gamma_2 (y,\ y), \qquad (17)$$

où

$$\gamma_1 = \frac{4}{h^2}\sin^2\frac{\pi}{2N} \geqslant \frac{8}{l^2}, \qquad \gamma_2 = \frac{4}{h^2}\cos^2\frac{\pi}{2N} < \frac{4}{h^2}.$$

En effet, soit $\mu_k(i)$ la fonction propre orthonormée du problème

$$(\mu_k)_{\bar{x}x} + \lambda_k\mu_k = 0, \quad 1 \leqslant i \leqslant N-1. \qquad \mu_k(0) = \mu_k(N) = 0. \tag{18}$$

On a noté au point 1, § 5, ch. I que la fonction de maille y_i remplissant les conditions du lemme peut être représentée sous forme de la somme

$$y_i = \sum_{k=1}^{N-1} c_k\mu_k(i), \quad c_k = (y, \mu_k). \tag{19}$$

A partir de (18) et (19) on tire

$$y_{\bar{x}x,\,i} = \sum_{k=1}^{N-1} c_k(\mu_k)_{\bar{x}x,\,i} = -\sum_{k=1}^{N-1}\lambda_k c_k\mu_k(i), \quad 1 \leqslant i \leqslant N-1.$$

En utilisant l'orthonormalité des fonctions propres μ_k, il vient

$$(y, y) = \sum_{k=1}^{N-1} c_k^2, \quad -(y_{\bar{x}x}, y) = \sum_{k=1}^{N-1}\lambda_k c_k^2. \tag{20}$$

En vertu de la première formule de différences de Green (12) on aura

$$-(y_{\bar{x}x}, y) = (y_{\bar{x}}^2, 1)_{\omega^+}. \tag{21}$$

Les valeurs propres λ_k du problème (18) ont été trouvées au point 1, § 5, ch. I:

$$\lambda_k = \frac{4}{h^2}\sin^2\frac{k\pi h}{2l} = \frac{4}{h^2}\sin^2\frac{k\pi}{2N}, \quad 1 \leqslant k \leqslant N-1,$$

avec

$$\gamma_1 = \min_k \lambda_k = \lambda_1 = \frac{4}{h^2}\sin^2\frac{\pi}{2N},$$

$$\gamma_2 = \max_k \lambda_k = \lambda_{N-1} = \frac{4}{h^2}\cos^2\frac{\pi}{2N}.$$

De là, ainsi que de (20), (21) on déduit l'estimation (17) du lemme 12

R e m a r q u e 1. Les estimations (17) sont précises au sens qu'elles passent à des égalités si en guise de y_i on prend $\mu_1(i)$ et $\mu_{N-1}(i)$. Notons que $\gamma_1 = 8/l^2$ si $h = l/2$, c'est-à-dire pour $N = 2$. Pour $N = 4$, on a $\gamma_1 = 32/(l^2(2+\sqrt{2})) > 8/l^2$.

R e m a r q u e 2. Si y_i ne devient nul que pour $i = 0$ ou $i = N$, alors dans (17) on a

$$\gamma_1 = \frac{4}{h^2}\sin^2\frac{\pi}{4N} \geqslant \frac{8}{l^2(2+\sqrt{2})}, \quad \gamma_2 = \frac{4}{h^2}\cos^2\frac{\pi}{4N} < \frac{4}{h^2}.$$

Si, par contre, y_i est une fonction de maille quelconque associée à $\bar{\omega}$, on a dans (17) $\gamma_1 = 0$ et $\gamma_2 = 4/h^2$. Pour démontrer ces assertions, il faut au lieu de (18) poser le problème correspondant de valeurs propres étudié au § 5, ch. I.

L'inégalité (17) peut être écrite sous la forme

$$\gamma_1 (y, y) \leqslant (-\Lambda y, y) \leqslant \gamma_2 (y, y), \tag{22}$$

si l'on introduit l'opérateur de différences Λ suivant la formule $\Lambda y_i = y_{\bar{x}x, i}$, $1 \leqslant i \leqslant N-1$ dans les fonctions y_i satisfaisant aux conditions $y_0 = y_N = 0$. Si la fonction de maille y_i ne devient nulle qu'à un bout du maillage $\bar{\omega}$, l'opérateur Λ doit être défini suivant les formules

$$\Lambda y_i = \begin{cases} y_{\bar{x}x, i}, & 1 \leqslant i \leqslant N-1, \\ -\frac{2}{h} y_{\bar{x}, i}, & i = N, \text{ si } y_0 = 0, \end{cases} \tag{23}$$

ou

$$\Lambda y_i = \begin{cases} \frac{2}{h} y_{x, i}, & i = 0, \\ y_{\bar{x}x, i}, & 1 \leqslant i \leqslant N-1, \quad \text{si } y_N = 0. \end{cases}$$

Compte tenu de ce que dans chacun de ces cas il s'ensuit de la première formule de différences de Green les égalités $(\Lambda y, y) = = (y_{\bar{x}}^2, 1)_{\omega^+}$, on obtient les inégalités (22), où γ_1 et γ_2 sont indiqués dans la remarque 2, tandis que y_i devient nul au bout correspondant du maillage $\bar{\omega}$.

Si y_i est une fonction de maille quelconque, l'opérateur Λ doit alors être défini ainsi:

$$\Lambda y_i = \begin{cases} \frac{2}{h} y_{x, 0}, & i = 0, \\ y_{\bar{x}x, i}, & 1 \leqslant i \leqslant N-1, \\ -\frac{2}{h} y_{\bar{x}, N}, & i = N. \end{cases}$$

Dans ce cas les inégalités (22) se vérifient également et

$$(-\Lambda y, y) = -(y_{\bar{x}x}, y) + y_{\bar{x}, N} y_N - y_{x, 0} y_0 = (y_{\bar{x}}^2, 1)_{\omega^+}.$$

Les constantes γ_1 et γ_2 sont indiquées dans la remarque 2.

Bref, on a trouvé les bornes pour les plus simples opérateurs de différences. Montrons maintenant que pour les opérateurs Λ introduits dans ce point se vérifie l'inégalité

$$| (-\Lambda u, v) | \leqslant (-\Lambda u, u)^{1/2} (-\Lambda v, v)^{1/2}. \tag{24}$$

Le principe d'obtention de l'inégalité (24) sera illustré sur l'exemple de l'opérateur $\Lambda y = y_{\bar{x}x}$. Introduisons l'espace $H(\omega)$ des fonc-

17–01162

tions de mailles, données sur ω avec produit scalaire $(u, v) = \sum_{i=1}^{N-1} u_i v_i h$, $u, v \in H(\omega)$. A l'opérateur de différences Λ correspond dans l'espace $H(\omega)$ l'opérateur linéaire A défini par l'égalité

$$Ay_i = -\Lambda \mathring{y}_i, \quad 1 \leqslant i \leqslant N-1,$$

où $y \in H(\omega)$, $y_i = \mathring{y}_i$ pour $1 \leqslant i \leqslant N-1$ et $\mathring{y}_0 = \mathring{y}_N = 0$. L'opérateur A constitue une application de $H(\omega)$ sur $H(\omega)$.

En vertu de l'égalité $(u, v) = (\mathring{u}, \mathring{v})$, on a $(Au, v) = -(\Lambda\mathring{u}, \mathring{v})$, où $\mathring{u}_0 = \mathring{u}_N = 0$, $\mathring{v}_0 = \mathring{v}_N = 0$. Il s'ensuit de (22) que $(Au, u) \geqslant \gamma_1(u, u)$, $\gamma_1 > 0$. Par conséquent, l'opérateur A est défini positif dans $H(\omega)$.

Montrons maintenant qu'il est autoadjoint dans $H(\omega)$. En effet, de la seconde formule de différences de Green (13) il s'ensuit

$$(Au, v) = -(\Lambda\mathring{u}, \mathring{v}) = -(\mathring{u}_{\bar{x}x}, \mathring{v}) = -(\mathring{u}, \mathring{v}_{\bar{x}x}) = (u, Av).$$

Vu que pour l'opérateur autoadjoint non négatif est satisfaite l'inégalité généralisée de Cauchy-Bouniakovski $|(Au, v)| \leqslant (Au, u)^{1/2} \times (Av, v)^{1/2}$, il vient de ce qui précède

$$|(-\Lambda\mathring{u}, \mathring{v})| \leqslant (-\Lambda\mathring{u}, \mathring{u})^{1/2} (-\Lambda\mathring{v}, \mathring{v})^{1/2},$$

ce qu'il fallait démontrer.

4. Estimations par le bas de quelques opérateurs de différences. Le lemme 12 a de fait établi les constantes de l'équivalence énergétique de l'opérateur unitaire E et de l'opérateur A correspondant à l'opérateur de différences $-\Lambda y = -y_{\bar{x}x}$ sur les fonctions qui deviennent nulles aux bouts du maillage $\bar{\omega}$, c'est-à-dire sur γ_1 et γ_2 de l'inégalité $\gamma_1 E \leqslant A \leqslant \gamma_2 E$.

Cherchons maintenant l'inégalité liant les opérateurs A et D, où $Dy_i = \rho_i y_i$, $1 \leqslant i \leqslant N-1$ et $\rho_i \geqslant 0$. A cette fin il nous faut déterminer la fonction de différences de Green de l'opérateur Λ.

Supposons qu'il s'agit de trouver sur le maillage $\bar{\omega}$ introduit plus haut la solution du problème de différences

$$\begin{aligned} &\Lambda v_i = v_{\bar{x}x,\,i} = -f_i, \quad 1 \leqslant i \leqslant N-1, \\ &v_0 = v_N = 0. \end{aligned} \tag{25}$$

La fonction de maille G_{ik} qui, une fois posé $k = 1, 2, \ldots, N-1$, satisfait aux conditions

$$\begin{aligned} &\Lambda G_{ik} = G_{\bar{x}x,\,ik} = -\frac{1}{h}\delta_{ik}, \quad 1 \leqslant i \leqslant N-1, \\ &G_{0k} = G_{Nk} = 0, \end{aligned}$$

où δ_{ik} est le symbole de Kronecker:

$$\delta_{ik} = \begin{cases} 1, & i = k; \\ 0, & i \neq k, \end{cases}$$

sera appelée *fonction de Green de l'opérateur de différences* Λ.

Fournissons les principales propriétés de la fonction de Green:

1) la fonction de Green est symétrique, $G_{ik} = G_{ki}$, de plus G_{ik} comme fonction de k, une fois posé $i = 1, 2, \ldots, N-1$, satisfait aux conditions

$$\Lambda G_{ik} = G_{\bar{x}x,\, ik} = -\frac{1}{h}\delta_{ik}, \quad 1 \leqslant k \leqslant N-1,$$
$$G_{i0} = G_{iN} = 0;$$

2) la fonction de Green est positive, $G_{ik} > 0$ pour $i, k \neq 0, N$;

3) pour toute fonction de maille y_i satisfaisant aux conditions $y_0 = y_N = 0$ est vraie la représentation

$$y_i = -\sum_{k=1}^{N-1} G_{ik}\Lambda y_k h, \tag{26}$$

de sorte que le problème (25) peut se représenter sous la forme

$$v_i = \sum_{k=1}^{N-1} G_{ik} f_k h, \quad 0 \leqslant i \leqslant N.$$

Cette assertion se démontre à l'aide de la seconde formule de différences de Green (13) et de la propriété 1).

L e m m e 13. *Soit* $\rho_i \geqslant 0$ *une fonction de maille donnée sur* ω *et non égale identiquement à zéro. Pour toute fonction de maille* y_i *donnée sur* $\bar{\omega}$ *et satisfaisant aux conditions* $y_0 = y_N = 0$, *se vérifie l'estimation*

$$\gamma_1 (\rho y, y) \leqslant (y_{\bar{x}}^2, 1)_{\omega^+}, \tag{27}$$

où $1/\gamma_1 = \max\limits_{1 \leqslant i \leqslant N-1} v_i$, *quant à* v_i *c'est la solution du problème aux limites*

$$\begin{aligned} &\Lambda v_i = v_{\bar{x}x,\, i} = -\rho_i, \quad 1 \leqslant i \leqslant N-1, \\ &v_0 = v_N = 0. \end{aligned} \tag{28}$$

En fait, posons $y_0 = y_N = 0$. En utilisant (26), il vient

$$(\rho y, y) = \sum_{i=1}^{N-1} \rho_i y_i^2 h = -\sum_{i=1}^{N-1} \rho_i y_i h \left(\sum_{k=1}^{N-1} G_{ik}\Lambda y_k h\right) =$$
$$= -\sum_{k=1}^{N-1} h\Lambda y_k \left(\sum_{i=1}^{N-1} \rho_i y_i G_{ik} h\right) = -(\Lambda y, w),$$

où on a posé $w_k = \sum_{i=1}^{N-1} \rho_i y_i G_{ik} h$, $0 \leqslant k \leqslant N$. En utilisant l'inégalité (24), on obtient de ce qui précède

$$(\rho y, y) \leqslant (-\Lambda y, y)^{1/2} (-\Lambda w, w)^{1/2}$$

ou, en vertu de (21),

$$(\rho y, y)^2 \leqslant (y_{\bar{x}}^2, 1)_{\omega^+} (-\Lambda w, w). \qquad (29)$$

Profitons de la propriété 1) de la fonction de Green G_{ik}. Il vient

$$-\Lambda w_k = -\sum_{i=1}^{N-1} h\rho_i y_i \Lambda G_{ik} = \sum_{i=1}^{N-1} \rho_i y_i \delta_{ik} = \rho_k y_k$$

et, par conséquent,

$$(-\Lambda w, w) = \sum_{k=1}^{N-1} h\rho_k y_k \Big(\sum_{i=1}^{N-1} h\rho_i y_i G_{ik}\Big) = \sum_{i=1}^{N-1} \sum_{k=1}^{N-1} a_{ik} y_i y_k,$$

où on a posé $a_{ik} = h^2 \rho_i \rho_k G_{ik}$, $1 \leqslant i,\ k \leqslant N-1$. En utilisant l'inégalité $2y_i y_k \leqslant y_i^2 + y_k^2$, de même que la symétrie et la positivité de la fonction de Green G_{ik}, on en tire

$$(-\Lambda w, w) \leqslant \sum_{i=1}^{N-1} 0{,}5 y_i^2 \sum_{k=1}^{N-1} a_{ik} + \sum_{k=1}^{N-1} 0{,}5 y_k^2 \sum_{i=1}^{N-1} a_{ki} =$$

$$= \sum_{i=1}^{N-1} y_i^2 \sum_{k=1}^{N-1} a_{ik} = \sum_{i=1}^{N-1} \rho_i y_i^2 h \Big(\sum_{k=1}^{N-1} \rho_k G_{ik} h\Big).$$

En vertu de la propriété 3) la solution du problème (28) s'écrit sous la forme

$$v_i = \sum_{k=1}^{N-1} \rho_k G_{ik} h > 0, \qquad 1 \leqslant i \leqslant N-1.$$

Donc

$$(-\Lambda w, w) = \sum_{i=1}^{N-1} \rho_i y_i^2 v_i h \leqslant \max_{1 \leqslant i \leqslant N-1} v_i (\rho y, y) = \frac{1}{\gamma_1} (\rho y, y).$$

De là et à partir de (29) s'ensuit l'estimation (27) du lemme.

Remarque 1. On peut montrer que la fonction $v_i = 0{,}5 x_i (l - x_i)$, où $x_i = ih \in [0, l]$, est la solution du problème (28) pour $\rho_i \equiv 1$. De là s'ensuit l'estimation

$$\gamma_1 (y, y) \leqslant (y_{\bar{x}}^2, 1)_{\omega^+}, \quad \gamma_1 = 8/l^2, \quad y_0 = y_N = 0. \qquad (30)$$

Remarque 2. Le lemme 13 peut être généralisé au cas quand y_i ne devient nul qu'à un bout du maillage $\bar{\omega}$. Par exemple, si $y_0 = 0$, on a dans (27) $1/\gamma_1 = \max\limits_{1 \leqslant i \leqslant N} v_i$, où v_i est la solution du

problème $\Lambda v_i = -\rho_i$, $1 \leqslant i \leqslant N$, $v_0 = 0$ à l'opérateur de différences Λ défini dans (23).

L e m m e 14. *Supposons que* $\rho_i \geqslant 0$, $d_i \geqslant 0$ *sont donnés sur* ω, *tandis que la fonction* $a_i \geqslant c_1 > 0$ *l'est sur* ω^+. *Pour toute fonction* y_i *donnée sur* $\bar{\omega}$ *et satisfaisant aux conditions* $y_0 = y_N = 0$ *se vérifie l'estimation*

$$\gamma_1 (\rho y,\ y) \leqslant (a y_{\bar{x}}^2,\ 1)_{\omega^+} + (dy,\ y),\quad 1/\gamma_1 = \max_{1 \leqslant i \leqslant N-1} v_i,$$

où v_i *est la solution du problème aux limites*

$$\Lambda v_i = (a v_{\bar{x}})_{x,\,i} - d_i v_i = -\rho_i,\quad 1 \leqslant i \leqslant N-1,\quad v_0 = v_N = 0.$$

R e m a r q u e 1. Si y_i ne devient nul qu'à l'un des bouts du maillage $\bar{\omega}$, par exemple pour $y_N = 0$, l'estimation

$$\gamma_1 (\rho y,\ y) \leqslant (a y_{\bar{x}}^2,\ 1)_{\omega^+} + (dy,\ y) + \varkappa_0 y_0^2, \tag{31}$$

où $1/\gamma_1 = \max\limits_{0 \leqslant i \leqslant N-1} v_i$ est vérifiée, tandis que la fonction v_i est la solution du problème

$$\Lambda v_i = -\rho_i,\quad 0 \leqslant i \leqslant N-1,\quad v_N = 0,$$

$$\Lambda y_i = \begin{cases} \frac{2}{h} (a_1 y_{x,\,0} - \varkappa_0 y_0) - d_0 y_0, & i = 0, \\ (a y_{\bar{x}})_{x,\,i} - d_i y_i, & 1 \leqslant i \leqslant N-1,\quad \varkappa_0 \geqslant 0. \end{cases} \tag{32}$$

R e m a r q u e 2. Pour une fonction de maille y_i quelconque donnée sur $\bar{\omega}$ on peut obtenir l'estimation

$$\gamma_1 (\rho y,\ y) \leqslant (a y_{\bar{x}}^2,\ 1)_{\omega^+} + (dy,\ y) + \varkappa_0 y_0^2 + \varkappa_1 y_N^2, \tag{33}$$

où $\varkappa_0 \geqslant 0$, $\varkappa_1 \geqslant 0$, $\varkappa_0 + \varkappa_1 + (d,\ 1) > 0$, tandis que les fonctions de mailles $\rho_i \geqslant 0$, $d_i \geqslant 0$ sont données sur $\bar{\omega}$. On a ici $1/\gamma_1 = \max\limits_{0 \leqslant i \leqslant N} v_i$, où v_i est la solution du problème aux limites

$$\Lambda v_i = -\rho_i,\quad 0 \leqslant i \leqslant N,$$

$$\Lambda y_i = \begin{cases} \frac{2}{h} (a_1 y_{x,\,0} - \varkappa_0 y_0) - d_0 y_0, & i = 0, \\ (a y_{\bar{x}})_{x,\,i} - d_i y_i, & 1 \leqslant i \leqslant N-1, \\ -\frac{2}{h} (a_N y_{\bar{x},\,N} + \varkappa_1 y_N) - d_N y_N, & i = N. \end{cases} \tag{34}$$

La démonstration du lemme 14 et des remarques 1 et 2 est conduite de la même façon que pour le lemme 13. On y utilise la fonction de Green des opérateurs de différences Λ déjà mentionnés, qui satisfait aux propriétés 1)-4) énumérées plus haut.

L e m m e 15. *Pour la fonction de maille y_i devenant nulle pour $i = N$ l'estimation*

$$y_0^2 \leqslant \text{th}(\varepsilon l)\left[\varepsilon(y, y) + \frac{1}{\varepsilon}(y_{\bar{x}}^2, 1)_{\omega^+}\right], \quad \varepsilon \geqslant 0, \tag{35}$$

est vraie. De façon analogue, l'estimation

$$y_N^2 \leqslant \text{th}(\varepsilon l)\left[\varepsilon(y, y) + \frac{1}{\varepsilon}(y_{\bar{x}}^2, 1)_{\omega^+}\right], \quad \varepsilon \geqslant 0,$$

est également vraie au cas où $y_0 = 0$. Pour une fonction de maille quelconque y_i donnée sur le maillage $\bar{\omega}$ on a l'estimation

$$y_0^2 + y_N^2 \leqslant \frac{8+\varepsilon^2 l^2}{\varepsilon l\sqrt{16+\varepsilon^2 l^2}}\left[\varepsilon(y, y) + \frac{1}{\varepsilon}(y_{\bar{x}}^2, 1)_{\omega^+}\right], \quad \varepsilon > 0. \tag{36}$$

Démontrons d'abord la justesse de l'estimation (35). A cette fin utilisons la remarque 1 du lemme 14. Posons dans (32) $a_i = 1/\varepsilon$, $d_i = \varepsilon$, $\varkappa_0 = 0$ et $\rho_0 = 2/h$, $\rho_i = 0$, $1 \leqslant i \leqslant N-1$. On obtient alors à partir de (31) l'estimation

$$y_0^2 \leqslant \max_{0\leqslant i\leqslant N-1} v_i\left[\varepsilon(y, y) + \frac{1}{\varepsilon}(y_{\bar{x}}^2, 1)_{\omega^+}\right],$$

où v_i est la solution du problème auxiliaire suivant:

$$\begin{gathered}\Lambda v_i = \frac{1}{\varepsilon}v_{\bar{x}x, i} - \varepsilon v_i = 0, \qquad 1 \leqslant i \leqslant N-1,\\ \Lambda v_0 = \frac{2}{\varepsilon h}v_{x,0} - \varepsilon v_0 = -\frac{2}{h}, \quad v_N = 0.\end{gathered} \tag{37}$$

Ecrivons (37) suivant les points

$$\begin{gathered}v_{i-1} - 2\alpha v_i + v_{i+1} = 0, \quad 1 \leqslant i \leqslant N-1,\\ v_1 - \alpha v_0 = -\varepsilon h, \quad v_N = 0,\end{gathered} \tag{38}$$

où $\alpha = 1 + 0{,}5\varepsilon^2 h^2 \geqslant 1$.

On obtient ainsi le problème aux limites sur l'équation aux différences du second ordre aux coefficients constants.

En se basant sur la théorie générale développée au point 1, § 4, ch. I ainsi que sur les propriétés du polynôme de Tchébychev (voir point 2 idem) on trouve que la fonction

$$v_i = \frac{\varepsilon h U_{N-i-1}(\alpha)}{T_N(\alpha)}, \quad 0 \leqslant i \leqslant N,$$

est la solution du problème (38). Ici

$$T_n(\alpha) = \text{ch}(n \operatorname{Arch}\alpha), \quad U_n(\alpha) = \frac{\text{sh}((n+1)\operatorname{Arch}\alpha)}{\text{sh}(\operatorname{Arch}\alpha)}, \quad |\alpha| \geqslant 1$$

sont des polynômes de Tchébychev du degré n de première et de seconde espèces.

Comme $\alpha \geqslant 1$, il s'ensuit que

$$\max_{0\leqslant i\leqslant N-1} v_i = v_0 = \frac{\varepsilon h U_{N-1}(\alpha)}{T_N(\alpha)}.$$

En résumé, on obtient l'estimation

$$y_0^2 \leqslant v_0\left[\varepsilon(y, y) + \frac{1}{\varepsilon}(y_{\bar{x}}^2, 1)_{\omega^+}\right]$$

pour la fonction de maille y_i satisfaisant à la condition $y_N = 0$. Cette estimation est vraie aux sens qu'elle passe à une égalité si en guise de y_i on prend la fonction v_i.

Apprécions maintenant v_0 par le haut pour tout h. Si l'on pose $\mathrm{ch}\, 2z = \alpha$, on a $z \geqslant 0$ et

$$\varepsilon h = 2\,\mathrm{sh}\, z, \quad N = l/n = \varepsilon l/(2\,\mathrm{sh}\, z),$$

$$T_N(\alpha) = \mathrm{ch}\, 2Nz = \mathrm{ch}\, w(z), \tag{39}$$

$$U_{N-1}(\alpha) = \frac{\mathrm{sh}\, 2Nz}{\mathrm{sh}\, 2z} = \frac{\mathrm{sh}\, w(z)}{2\,\mathrm{sh}\, z\,\mathrm{ch}\, z}, \quad w(z) = \frac{\varepsilon l z}{\mathrm{sh}\, z}.$$

Donc

$$v_0 = \frac{\mathrm{sh}\, w(z)}{\mathrm{ch}\, z\,\mathrm{ch}\, w(z)}.$$

Comme pour un ε fixé

$$\frac{dw}{dz} = \frac{\varepsilon l(\mathrm{sh}\, z - z\,\mathrm{ch}\, z)}{\mathrm{sh}^2 z} \leqslant 0,$$

il vient que

$$\frac{dv_0}{dz} = \frac{\mathrm{ch}\, z \dfrac{dw}{dz} - \mathrm{sh}\, z\,\mathrm{sh}\, w\,\mathrm{ch}\, w}{\mathrm{ch}^2 z\,\mathrm{ch}^2 w} \leqslant 0.$$

Par conséquent, pour $z = 0$ v_0 est maximal. On a ainsi l'estimation $v_0 \leqslant \mathrm{th}(\varepsilon l)$. L'inégalité (35) est démontrée.

Soit maintenant une fonction de maille y_i quelconque. A partir de la remarque 2 du lemme 14 au cas où $\alpha_i \equiv 1/\varepsilon$, $d_i \equiv \varepsilon$, $\varkappa_0 = \varkappa_1 = 0$, $\rho_0 = \rho_N = 2/h$, $\rho_i = 0$ avec $1 \leqslant i \leqslant N-1$, on déduit l'estimation

$$y_0^2 + y_N^2 \leqslant \max_{0 \leqslant i \leqslant N} v_i \left[\varepsilon(y, y) + \frac{1}{\varepsilon}(y_{\bar{x}}^2, 1)_{\omega^+}\right],$$

où v_i est la solution du problème aux limites

$$\frac{1}{\varepsilon} v_{\bar{x}x,\, i} - \varepsilon v_i = 0, \quad 1 \leqslant i \leqslant N-1,$$

$$\frac{2}{\varepsilon h} v_{x,\, 0} - \varepsilon v_0 = -\frac{2}{h}, \quad -\frac{2}{\varepsilon h} v_{\bar{x},\, N} - \varepsilon v_N = -\frac{2}{h}. \tag{40}$$

La solution du problème (40) est la fonction

$$v_i = \frac{\varepsilon h\,[T_{N-i}(\alpha) + T_i(\alpha)]}{(\alpha^2 - 1)\, U_{N-1}(\alpha)}, \quad 0 \leqslant i \leqslant N,$$

où α est défini plus haut.

De là on obtient que

$$\max_{0 \leqslant i \leqslant N} v_i = v_0 = v_N = \frac{\varepsilon h\,(1 + T_N(\alpha))}{(\alpha^2 - 1)\, U_{N-1}(\alpha)}. \tag{41}$$

Apprécions cette expression par le haut pour un h quelconque. En utilisant (39), il vient

$$v_0 = \frac{1 + \mathrm{ch}\, w(z)}{\mathrm{ch}\, z\,\mathrm{sh}\, w(z)} = \frac{\mathrm{ch}\, \frac{1}{2} w(z)}{\mathrm{ch}\, z\,\mathrm{sh}\, \frac{1}{2} w(z)} \leqslant \frac{\mathrm{ch}\, \frac{1}{2} w(z)}{\mathrm{sh}\, \frac{1}{2} w(z)} = \varphi(z).$$

Etant donné que

$$\frac{d\varphi}{dz} = -\frac{1}{\mathrm{sh}^2 0{,}5w} \frac{\partial w}{\partial z} > 0,$$

la fonction $\varphi(z)$ est maximale pour $z = z_0$ maximal qu'on tire de la relation $\mathrm{ch}\, 2z_0 = 1 + \varepsilon^2 l^2/8$ ($h \leqslant l/2$). A partir de (39) on obtient que $w(z_0) = 4z_0$. Donc

$$\varphi(z_0) = \frac{\mathrm{ch}\, 2z_0}{\mathrm{sh}\, 2z_0} = \frac{1+\varepsilon^2 l^2/8}{\sqrt{\varepsilon^2 l^2/8 + \varepsilon^4 l^4/64}} = \frac{8+\varepsilon^2 l^2}{\varepsilon l \sqrt{16+\varepsilon^2 l^2}}.$$

On a ainsi obtenu l'estimation (36).

Les lemmes 13 et 14 peuvent être généralisés sans peine au cas d'un maillage irrégulier quelconque $\bar{\omega}$. Dans ce cas on utilise pour les produits scalaires les notations (4), (6), quant aux opérateurs de différences Λ, ils sont remplacés par des opérateurs adéquats sur le maillage irrégulier.

L e m m e 16. *Supposons que* $\rho_i \geqslant 0$, $d_i \geqslant 0$ *sont donnés sur un maillage irrégulier quelconque* $\bar{\omega}$, $\rho_i \not\equiv 0$ *et* $a_i \geqslant c_i > 0$ *étant donnés sur* ω^+. *Soient* $\varkappa_0 \geqslant 0$, $\varkappa_1 \geqslant 0$ *des nombres quelconques avec la condition* $\varkappa_0 + \varkappa_1 + (d, 1) > 0$ *satisfaite. Pour toute fonction de maille* y_i *donnée sur* $\bar{\omega}$ *l'inégalité* (33), *où* $1/\gamma_1 = \max\limits_{0 \leqslant i \leqslant N} v_i$, *se vérifie,* v_i *étant la solution du problème* $\Lambda v_i = -\rho_i$, $0 \leqslant i \leqslant N$. *L'opérateur* Λ *est défini ici par les formules*

$$\Lambda y_i = \begin{cases} \frac{1}{\hbar_0}(a_1 y_{x,0} - \varkappa_0 y_0) - d_0 y_0, & i = 0, \\ (a y_{\bar{x}})_{\hat{x},i} - d_i y_i, & 1 \leqslant i \leqslant N-1, \\ -\frac{1}{\hbar_N}(a_N y_{\bar{x},N} + \varkappa_1 y_N) - d_N y_N, & i = N. \end{cases} \qquad (42)$$

Le lemme 16 se démontre de la même façon que les lemmes précédents.

R e m a r q u e 1. Si $a_i \equiv 1$, $d_i \equiv 0$, $\rho_i \equiv 1$, l'inégalité (33) prend la forme

$$\gamma_1 (y, y) \leqslant (y_{\bar{x}}^2, 1)_{\omega^+} + \varkappa_0 y_0^2 + \varkappa_1 y_N^2, \qquad (43)$$

où

$$\gamma_1 = \frac{8(\varkappa_0 + \varkappa_1 + l\varkappa_0\varkappa_1)^2}{l(2+l\varkappa_0)(2+l\varkappa_1)(2\varkappa_0 + 2\varkappa_1 + l\varkappa_0\varkappa_1)}.$$

Si de plus $y_0 = y_N = 0$, l'inégalité (43) passe alors à l'inégalité (30). Si y_i ne devient nul qu'à un bout, par exemple, pour $i = N$, alors, en posant dans (43) $y_N = 0$ et en passant à la limite pour $\varkappa_1 \to \infty$, on obtient l'estimation

$$\gamma_1 (y, y) \leqslant (y_{\bar{x}}^2, 1)_{\omega^+} + \varkappa_0 y_0^2, \quad \gamma_1 = \frac{8(1+l\varkappa_0)^2}{l^2(2+l\varkappa_0)^2}.$$

R e m a r q u e 2. De la définition donnée dans (42) de l'opérateur de différences Λ et de la première formule de différences de

Green il s'ensuit que

$$(-\Lambda y,\ y)=(ay_{\bar{x}}^2,\ 1)_{\omega^+}+(dy,\ y)+\varkappa_0 y_0^2+\varkappa_1 y_N^2.$$

Aussi l'inégalité (33) du lemme 16 peut-elle être écrite sous forme

$$\gamma_1(\rho y,\ y)\leqslant -(\Lambda y,\ y).$$

Passons à la déduction de l'estimation (43). Cherchons la solution du problème $\Lambda v_i=-\rho_i$, $0\leqslant i\leqslant N$, avec les hypothèses mentionnées dans la remarque 1. On a le problème aux limites au sens des différences finies

$$v_{\bar{x}x,\ i}=-1,\qquad 1\leqslant i\leqslant N-1, \tag{44}$$

$$v_{x,\ 0}=\varkappa_0 v_0-\hbar_0,\qquad i=0, \tag{45}$$

$$-v_{\bar{x},\ N}=\varkappa_1 v_N-\hbar_N,\qquad i=N. \tag{46}$$

Multiplions l'équation (44) par $\hbar_i$ et sommons en i de j à $N-1$, compte tenu de la condition aux limites (46). Il vient

$$\sum_{i=j}^{N-1} v_{\bar{x}x,\ i}\hbar_i=\sum_{i=j}^{N-1}(v_{\bar{x},\ i+1}-v_{\bar{x},\ i})=v_{\bar{x},\ N}-v_{\bar{x},\ j}=$$

$$=-\varkappa_1 v_N+\hbar_N-v_{\bar{x},\ j}=-\sum_{i=j}^{N-1}\hbar_i=x_j-0{,}5h_j-l+\hbar_N.$$

De là il s'ensuit que

$$v_{\bar{x},\ j}=l-\varkappa_1 v_N+0{,}5h_j-x_j,\qquad 1\leqslant j\leqslant N. \tag{47}$$

En posant dans (47) $j=1$ et compte tenu de l'égalité $\hbar_0=0{,}5h_1$, $v_{\bar{x},\ 1}=v_{x,\ 0}=$ $=\varkappa_0 v_0-\hbar_0$, on obtient la relation entre v_0 et v_N

$$\varkappa_0 v_0+\varkappa_1 v_N=l. \tag{48}$$

En multipliant (47) par h_j et en sommant en j de 1 à i, on obtient

$$\sum_{j=1}^{i} v_{\bar{x},\ j}h_j=v_i-v_0=(l-\varkappa_1 v_N)\sum_{j=1}^{i} h_j-\sum_{j=1}^{i}(x_j-0{,}5h_j)\,h_j.$$

Vu que $h_j=x_j-x_{j-1}$, $x_j-0{,}5h_j=0{,}5\,(x_j+x_{j-1})$, il vient

$$\sum_{j=1}^{i} h_j=x_i,\quad \sum_{j=1}^{i}(x_j-0{,}5h_j)\,h_j=0{,}5\sum_{j=1}^{i}(x_j^2-x_{j-1}^2)=0{,}5x_i^2.$$

On a donc

$$v_i=v_0+x_i\,(l-\varkappa_1 v_N)-0{,}5x_i^2=$$
$$=v_0+0{,}5\,(l-\varkappa_1 v_N)^2-0{,}5\,(x_i-l+\varkappa_1 v_N)^2,\ 0\leqslant i\leqslant N. \tag{49}$$

En posant ici $i=N$, on obtient la seconde relation entre v_0 et v_N

$$v_N=v_0+l\,(l-\varkappa_1 v_N)-0{,}5l^2. \tag{50}$$

De (48), (50) on tire

$$v_0=\frac{l\,(2+l\varkappa_1)}{2\,(\varkappa_0+\varkappa_1+\varkappa_0\varkappa_1 l)},\qquad v_N=\frac{l\,(2+l\varkappa_0)}{2\,(\varkappa_0+\varkappa_1+\varkappa_0\varkappa_1 l)}. \tag{51}$$

Vu que $0\leqslant l-\varkappa_1 v_N<l$, à partir de (49), (51) il s'ensuit que

$$\max_{0\leqslant i\leqslant N} v_i\leqslant v_0+0{,}5\,(l-\varkappa_1 v_N)^2=$$

$$=\frac{l\,(2+l\varkappa_0)\,(2+l\varkappa_1)\,(2\varkappa_0+2\varkappa_1+l\varkappa_0\varkappa_1)}{8\,(\varkappa_0+\varkappa_1+l\varkappa_0\varkappa_1)^2}.$$

De là et à partir du lemme 16 s'ensuit l'estimation (43). Si $y_0 = y_N = 0$, alors en posant dans (33) $a_i \equiv 1$, $d_i \equiv 0$, $\rho_i \equiv 1$ et en passant dans (43) à la limite pour $\varkappa_0 \to \infty$ et $\varkappa_1 \to \infty$, on obtient l'estimation (30) avec $\gamma_1 = 8/l^2$.

5. Appréciation par le haut d'opérateurs de différences. Cherchons maintenant l'estimation par le haut de quelques opérateurs de différences.

Lemme 17. *Pour une fonction de maille quelconque y_i donnée sur un maillage irrégulier $\bar{\omega}$ se vérifie l'estimation*

$$(ay_{\bar{x}}^2,\ 1)_{\omega^+} \leqslant \gamma_2 (y,\ y), \tag{52}$$

où

$$\gamma_2 = \max\left[\frac{4a_1}{h_1^2},\ \frac{4a_N}{h_N^2},\quad \max_{1\leqslant i\leqslant N-1} \frac{2}{\hbar_i}\left(\frac{a_i}{h_i} + \frac{a_{i+1}}{h_{i+1}}\right)\right].$$

Si le maillage est régulier, alors

$$\gamma_2 = \frac{4}{h^2} \max\left[a_1,\ a_N,\quad \max_{1\leqslant i\leqslant N-1}\left(\frac{a_i + a_{i+1}}{2}\right)\right].$$

Si $y_0 = y_N = 0$, $\gamma_2 = \max\limits_{1\leqslant i\leqslant N-1} \dfrac{2}{\hbar_i}\left(\dfrac{a_i}{h_i} + \dfrac{a_{i+1}}{h_{i+1}}\right)$.

On a en effet

$$(ay_{\bar{x}}^2,\ 1)_{\omega^+} = \sum_{i=1}^{N} \frac{a_i (y_i - y_{i-1})^2}{h_i} =$$

$$= \sum_{i=1}^{N} \frac{a_i}{h_i} y_i^2 + \sum_{i=0}^{N-1} \frac{a_{i+1}}{h_{i+1}} y_i^2 - 2\sum_{i=1}^{N} \frac{a_i}{h_i} y_i y_{i-1}.$$

En utilisant l'inégalité $2y_iy_{i-1} \leqslant y_i^2 + y_{i-1}^2$, on obtient pour $a_i > 0$ que

$$(ay_{\bar{x}}^2,\ 1)_{\omega^+} \leqslant \sum_{i=1}^{N} \frac{2a_i}{h_i} y_i^2 + \sum_{i=0}^{N-1} \frac{2a_{i+1}}{h_{i+1}} y_i^2 =$$

$$= \frac{2a_1}{h_1\hbar_0} y_0^2\hbar_0 + \frac{2a_N}{h_N\hbar_N} y_N^2\hbar_N + \sum_{i=1}^{N-1} \frac{2}{\hbar_i}\left(\frac{a_i}{h_i} + \frac{a_{i+1}}{h_{i+1}}\right) y_i^2\hbar_i.$$

Vu que $\hbar_0 = 0{,}5h_1$, $\hbar_N = 0{,}5h_N$ et $(y,\ y) = \sum\limits_{i=0}^{N} y_i^2\hbar_i$, il en suit l'estimation (52) avec la valeur indiquée pour γ_2. Le lemme 17 est démontré.

Lemme 18. *Soient $a_i > 0$, $b_i \geqslant 0$, tandis que σ_0 et σ_1 sont non négatifs avec $(b,\ 1) + \sigma_0 + \sigma_1 \neq 0$. Pour une fonction de maille*

quelconque y_i *donnée sur un maillage irrégulier* $\bar{\omega}$ *se vérifie l'estimation*

$$(ay_{\bar{x}}^2, 1)_{\omega^+} + (by, y) + \sigma_0 y_0^2 + \sigma_1 y_N^2 \leqslant \bar{\gamma}_2 (y, y), \tag{53}$$

où $\bar{\gamma}_2 = \gamma_2 + (1 + \gamma_2) \max\limits_{0 \leqslant i \leqslant N} v_i$, γ_2 *est défini dans le lemme* 17 *et* v *la solution du problème aux limites*

$$\begin{gathered}(av_{\bar{x}})_{\hat{x}, i} - v_i = -b_i, \quad 1 \leqslant i \leqslant N-1, \\ \frac{a_1}{\hbar_0} v_{x, 0} - v_0 = -b_0 - \frac{\sigma_0}{\hbar_0}, \quad i = 0, \\ -\frac{a_N}{\hbar_N} v_{\bar{x}, N} - v_N = -b_N - \frac{\sigma_1}{\hbar_N}, \quad i = N.\end{gathered} \tag{54}$$

En effet, à partir du lemme 16 avec $\rho_i = b_i$ pour $1 \leqslant i \leqslant N-1$, $\rho_0 = b_0 + \sigma_0/\hbar_0$, $\rho_N = b_N + \sigma_1/\hbar_N$ et $\varkappa_0 = \varkappa_1 = 0$, $d_i \equiv 1$, on obtient l'estimation

$$(by, y) + \sigma_0 y_0^2 + \sigma_1 y_N^2 = (\rho y, y) \leqslant \max_{0 \leqslant i \leqslant N} v_i [(ay_{\bar{x}}^2, 1)_{\omega^+} + (y, y)],$$

où v_i est la solution du problème auxiliaire (54). En utilisant le lemme 17, on a

$$(ay_{\bar{x}}^2, 1)_{\omega^+} + (by, y) + \sigma_0 y_0^2 + \sigma_1 y_N^2 \leqslant (1 + c)\,(ay_{\bar{x}}^2, 1)_{\omega^+} + \\ + c\,(y, y) \leqslant [\gamma_2 + (1 + \gamma_2)\,c]\,(y, y), \quad c = \max_{0 \leqslant i \leqslant N} v_i.$$

Le lemme 18 est démontré.

6. Schémas aux différences en tant que équations opératorielles dans des espaces abstraits. Après remplacement des dérivées entrant dans les équations différentielles et les conditions aux limites par des rapports incrémentiels sur maillage $\bar{\omega}$ choisi, on obtient un schéma aux différences. Les équations aux différences reliant les valeurs cherchées de la fonction de maille aux nœuds $\bar{\omega}$ constituent un système d'équations algébriques. Ce système est linéaire si le problème initial était linéaire.

Le schéma aux différences est défini par un opérateur de différences, fixant la structure des équations aux différences aux nœuds du maillage où l'on recherche la solution du problème, et par des conditions aux limites imposées aux nœuds frontières. L'opérateur de différences agit dans l'espace des fonctions de mailles associées à $\bar{\omega}$.

Voyons un exemple. Supposons qu'il s'agit d'obtenir la solution du problème

$$\begin{gathered}u'' = -\varphi(x), \quad 0 < x < l, \\ u'(0) = \varkappa_0 u(0) - \mu_1, \quad u(l) = \mu_2, \quad \varkappa_0 \geqslant 0\end{gathered} \tag{55}$$

sur le tronçon $0 \leqslant x \leqslant l$.

Sur un maillage régulier $\bar{\omega} = \{x_i = ih,\ i = 0, 1, \ldots, N,\ hN = l\}$ le problème (55) sera mis en accord avec le schéma aux différences

$$\begin{aligned} &\Lambda y_i = y_{\bar{x}x,\, i} = -\varphi_i, \quad 1 \leqslant i \leqslant N-1, \\ &\Lambda y_0 = \frac{2}{h}(y_{x,\,0} - \varkappa_0 y_0) = -\left(\varphi_0 + \frac{2}{h}\mu_1\right), \\ &y_N = \mu_2. \end{aligned} \tag{56}$$

L'opérateur de différences Λ est défini sur un $(N+1)$-ème ensemble des fonctions de mailles données sur $\bar{\omega}$ et constitue son application sur le N-ème ensemble des fonctions données sur $\omega^- = \{x_i \in \bar{\omega},\ i = 0, 1, \ldots, N-1\}$. On voit que le domaine de définition et le domaine des valeurs de l'opérateur Λ ne coïncident pas.

Voyons maintenant l'espace $H(\omega^-)$ de fonctions de mailles données sur ω^-. Définissons le produit scalaire dans $H(\omega^-)$ comme on l'a fait dans l'exemple 1 du point 1, § 2:

$$(u, v) = \sum_{i=1}^{N-1} u_i v_i h + 0{,}5 h u_0 v_0, \quad u,\ v \in H(\omega^-).$$

Définissons maintenant l'opérateur linéaire A de la façon suivante: $Ay_i = -\Lambda \overset{\circ}{y}_i$, $0 \leqslant i \leqslant N-1$, où $y \in H(\omega^-)$, $\overset{\circ}{y}_i = y_i$ pour $0 \leqslant i \leqslant N-1$ et $\overset{\circ}{y}_N = 0$. En utilisant cette définition, donnons la transcription détaillée de l'opérateur A:

$$Ay_i = \begin{cases} -\dfrac{2}{h}(y_{x,\,0} - \varkappa_0 y_0), & i = 0, \\ -y_{\bar{x}x,\, i}, & 1 \leqslant i \leqslant N-2, \\ \dfrac{1}{h^2}(2y_{N-1} - y_{N-2}), & i = N-1. \end{cases} \tag{57}$$

L'opérateur A constitue l'application $H(\omega^-)$ sur $H(\omega^-)$ et est linéaire.

Transformons le schéma aux différences (2). Compte tenu de la condition $y_N = \mu_2$, écrivons (56) sous la forme

$$\begin{aligned} &-\frac{2}{h}(y_{x,\,0} - \varkappa_0 y_0) = f_0 = \left(\varphi_0 + \frac{2}{h}\mu_1\right), \\ &-y_{\bar{x}x,\, i} = f_i = \varphi_i, \quad 1 \leqslant i \leqslant N-2, \\ &\frac{1}{h^2}(2y_{N-1} - y_{N-2}) = f_{N-1} = \left(\varphi_{N-1} + \frac{1}{h^2}\mu_2\right). \end{aligned} \tag{58}$$

En comparant (57) à (58), on trouve que le schéma aux différences (56) s'écrit sous forme d'une équation opératorielle de première espèce

$$Ay = f, \tag{59}$$

où y est l'élément inconnu, f l'élément donné de l'espace $H(\omega^-)$, tandis que A, l'opérateur agissant dans $H(\omega^-)$, est défini plus haut.

Indiquons les principales propriétés de l'opérateur A.

L'opérateur A est autoadjoint dans $H(\omega^-)$, c'est-à-dire que

$$(Au, v) = (u, Av), \quad u, v \in H(\omega^-).$$

En effet, $(Au, v) = -(\Lambda\dot{u}, \dot{v})$ avec $\dot{u}_N = \dot{v}_N = 0$. En utilisant la seconde formule de différences de Green (13), on obtient

$$(\Lambda\mathring{u}, \mathring{v}) = \sum_{i=1}^{N-1} \mathring{u}_{\bar{x}x,i}\mathring{v}_i h + (\mathring{u}_{x,0} - \varkappa_0\mathring{u}_0)\mathring{v}_0 =$$

$$= \sum_{i=1}^{N-1} \mathring{u}_i\mathring{v}_{\bar{x}x,i}h + (\mathring{u}_{\bar{x}}\mathring{v} - \mathring{v}_{\bar{x}}\mathring{u})_N - (\mathring{u}_x\mathring{v} - \mathring{v}_x\mathring{u})_0 +$$

$$+ (\mathring{u}_x\mathring{v} - \varkappa_0\mathring{u}\mathring{v})_0 = \sum_{i=1}^{N-1} \mathring{u}_i\mathring{v}_{\bar{x}x,i}h + (\mathring{v}_x\mathring{u} - \varkappa_0\mathring{v}\mathring{u})_0 = (\mathring{u}, \Lambda\mathring{v}).$$

La proposition est démontrée.

L'opérateur A est défini positif, c'est-à-dire

$$(Au, u) \geqslant \gamma_1(u, u), \quad u \in H(\omega^-),$$

où $\gamma_1 = \frac{8(1+l\varkappa_0)^2}{l^2(2+l\varkappa_0)^2} \geqslant \frac{2}{l^2} > 0$. Cette proposition s'ensuit des remarques 1 et 2 associées au lemme 16. L'opérateur A, en vertu du lemme 10, possède un opérateur inverse A^{-1} borné. Aussi l'équation (59) possède-t-elle une solution qui est unique.

On a pour l'opérateur A l'estimation par le haut

$$(Au, u) \leqslant \gamma_2(u, u), \quad u \in H(\omega^-),$$

où $\gamma = \frac{4}{h^2}\left(1 + \varkappa_0\frac{h}{2}\right)$, vu que $y_N = 0$ et

$$(Ay, y) = (y_{\bar{x}}^2, 1)_{\omega^+} + \varkappa_0 y_0^2,$$

$$y_0^2 \leqslant \frac{2}{h}(y, y), \quad (y_{\bar{x}}^2, 1)_{\omega^+} \leqslant \frac{4}{h^2}.$$

Cette dernière inégalité s'ensuit du lemme 17.

En guise de second exemple, examinons sur un maillage irrégulier $\bar{\omega} = \{x_i \in [0, l], x_i = x_{i-1} + h_i, 1 \leqslant i \leqslant N, x_0 = 0, x_N = l\}$ le schéma aux différences

$$\Lambda y_i = (ay_{\bar{x}})_{\hat{x},i} - d_i y_i = -\varphi_i, \quad 1 \leqslant i \leqslant N-1,$$

$$\Lambda y_0 = \frac{1}{\hbar_0}(a_1 y_{x,0} - \varkappa_0 y_0) - d_0 y_0 = -\left(\varphi_0 + \frac{1}{\hbar_0}\mu_1\right), \quad i = 0, \qquad (60)$$

$$\Lambda y_N = -\frac{1}{\hbar_N}(a_N y_{\bar{x},N} + \varkappa_1 y_N) - d_N y_N = -\left(\varphi_N + \frac{1}{\hbar_N}\mu_2\right), \quad i = N.$$

Le schéma (60) constitue une approximation du troisième problème aux limites sur l'équation aux coefficients variables

$$(ku')' - qu = -\varphi(x), \quad 0 < x < l,$$
$$ku' = \varkappa_0 u - \mu_1, \quad x = 0,$$
$$-ku' = \varkappa_1 u - \mu_2, \quad x = l$$

au cas d'un choix adéquat des coefficients a_i et d_i, par exemple, pour $a_i = k(x_i - 0{,}5h_i)$ et $d_i = q(x_i)$.

Si dans un espace $H(\omega^-)$ des fonctions de mailles données sur $\overline{\omega}$ avec produits scalaires

$$(u, v) = \sum_{i=0}^{N} u_i v_i \hbar_i, \quad \hbar_0 = 0{,}5h_1, \quad \hbar_N = 0{,}5h_N,$$

on définit l'opérateur $A = -\Lambda$ et la fonction de maille $f_i = \varphi_i$, $1 \leqslant i \leqslant N-1$, $f_0 = \varphi_0 + \mu_1/\hbar_0$, $f_N = \varphi_N + \mu_2/\hbar_N$, on peut alors transcrire le schéma aux différences (60) sous forme d'équation opératorielle (59).

Le fait que l'opérateur A, application de $H(\overline{\omega})$ sur $H(\overline{\omega})$, est autoadjoint s'ensuit de la seconde formule de différences de Green.

Si les conditions $a_i \geqslant c_1 > 0$, $d_i \geqslant 0$, $\varkappa_0 \geqslant 0$, $\varkappa_1 \geqslant 0$, $\varkappa_0 + \varkappa_1 + (d, 1) > 0$ sont remplies, l'opérateur A est défini positif dans $H(\overline{\omega})$ et l'estimation $(Au, u) \geqslant \gamma_1 (u, u)$, $1/\gamma_1 = \max\limits_{0 \leqslant i \leqslant N} v_i$, où v_i est la solution du problème $\Lambda v_i = -1$, $0 \leqslant i \leqslant N$, est vérifiée. Notons que la positivité de v_i s'ensuit du principe du maximum se justifiant pour l'opérateur Λ dans les conditions indiquées.

Si $d_i \equiv 0$, on est en mesure d'obtenir une estimation grossière de γ_1 de la façon suivante. A partir de la première formule de différences de Green on obtient

$$(Ay, y) = (-\Lambda y, y) = (ay_{\bar{x}}^2, 1)_{\omega^+} + \varkappa_0 y_0^2 + \varkappa_1 y_1^2.$$

En vertu des conditions $a_i \geqslant c_1 > 0$, $1 \leqslant i \leqslant N$, on obtient

$$(Ay, y) \geqslant c_1 [(y_{\bar{x}}^2, 1)_{\omega^+} + \bar{\varkappa}_0 y_0^2 + \bar{\varkappa}_1 y_1^2],$$

où $c_1 \bar{\varkappa}_0 = \varkappa_0$, $c_1 \bar{\varkappa}_1 = \varkappa_1$. Comme $\varkappa_0 + \varkappa_1 > 0$, de la remarque 1 du lemme 16 on obtient l'estimation

$$(y_{\bar{x}}^2, 1)_{\omega^+} + \bar{\varkappa}_0 y_0^2 + \varkappa_1 y_1^2 \geqslant \bar{\gamma}_1 (y, y),$$

où

$$\bar{\gamma}_1 = \frac{8(\bar{\varkappa}_0 + \bar{\varkappa}_1 + l\bar{\varkappa}_0\bar{\varkappa}_1)^2}{l(2 + l\bar{\varkappa}_0)(2 + l\bar{\varkappa}_1)(2\bar{\varkappa}_0 + 2\bar{\varkappa}_1 + l\bar{\varkappa}_0\bar{\varkappa}_1)}.$$

En y portant $\bar{\varkappa}_0$ et $\bar{\varkappa}_1$, on trouve que $(Au, u) \geqslant \gamma_1 (u, u)$, où

$$\gamma_1 = c_1 \bar{\gamma}_1 = \frac{8c_1(c_1\varkappa_0 + c_1\varkappa_1 + l\varkappa_0\varkappa_1)^2}{l(2c_1 + l\varkappa_0)(2c_1 + l\varkappa_1)(2c_1\varkappa_0 + 2c_1\varkappa_1 + l\varkappa_0\varkappa_1)}.$$

Pour l'opérateur A a lieu l'estimation par le haut $(Au, u) \leqslant \gamma_2 (u, u)$, où γ_2 est défini dans le lemme 18, car

$$(Ay, y) = (ay_{\bar{x}}^2, 1)_{\omega^+} + (dy^2, 1) + \varkappa_0 y_0^2 + \varkappa_1 y_N^2.$$

Dans l'exemple étudié l'opérateur A et l'opérateur de différences Λ sont définis dans un même espace de fonctions de mailles $H(\bar{\omega})$ et ne diffèrent que par le signe. À la différence du premier exemple, les seconds membres du schéma aux différences (60) et de l'équation opératorielle (59) coïncident.

On s'est ici limité à des exemples les plus simples. Au point suivant les schémas aux différences approximant les problèmes aux limites elliptiques dans un espace à plusieurs dimensions seront réduits de façon analogue à des équations opératorielles dans des espaces hilbertiens appropriés de dimensions finies des fonctions de mailles. On y étudiera également les principales propriétés de tels opérateurs.

Les exemples cités montrent que les schémas aux différences peuvent être assimilés à des équations opératorielles dont les opérateurs sont définis dans un espace linéaire normé de dimensions finies. Ces opérateurs se caractérisent par le fait qu'ils constituent une application de tout l'espace en eux-mêmes.

7. Schémas aux différences pour des équations elliptiques à coefficients constants. Soit $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha, \alpha = 1, 2\}$ un rectangle, $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leqslant i \leqslant N_1, 0 \leqslant j \leqslant N_2, h_\alpha N_\alpha = l_\alpha, \alpha = 1, 2\}$ un maillage dans $\bar{G}$, γ un ensemble de nœuds frontières du maillage $\bar{\omega}$. Le maillage est régulier dans chaque direction x_α de pas h_α. Désignons par ω l'ensemble des nœuds intérieurs du maillage. Introduisons l'espace des fonctions de mailles $H = H(\omega)$ données sur ω. Définissons dans H le produit scalaire

$$(u, v) = \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} u(i, j) v(i, j) h_1 h_2.$$

Etudions le problème de différences de Dirichlet pour l'équation de Poisson associée au maillage $\bar{\omega}$

$$\Lambda y = \sum_{\alpha=1}^{2} \Lambda_\alpha y = -\varphi(x), \quad x \in \omega, \qquad y(x) = g(x), \quad x \in \gamma, \tag{61}$$

où $\Lambda_\alpha y = y_{\bar{x}_\alpha x_\alpha}$, $\alpha = 1, 2$.

Le schéma aux différences (61) peut être transcrit sous forme d'équation opératorielle (59). A cette fin définissons l'opérateur A suivant la formule $Ay = -\Lambda \mathring{y}$, $x \in \omega$, où $y \in H$, $\mathring{y} \in \mathring{H}$ et $y(x) = \mathring{y}(x)$ pour $x \in \omega$. $\mathring{H}$ est ici un ensemble des fonctions de mailles

données sur $\bar{\omega}$ et devenant nulles sur γ. Le second membre f de l'équation (59) ne diffère de celui de φ du schéma aux différences (61) qu'aux nœuds frontières

$$f = \varphi + \varphi_1/h_1^2 + \varphi_2/h_2^2,$$

où

$$\varphi_1(x) = \begin{cases} g(0,\ x_2), & x_1 = h_1, \\ 0, & 2h_1 \leqslant x_1 \leqslant l_1 - 2h_1, \\ g(l_1,\ x_2), & x_1 = l_1 - h_1, \end{cases}$$

$$\varphi_2(x) = \begin{cases} g(x_1,\ 0), & x_2 = h_2, \\ 0, & 2h_2 \leqslant x_2 \leqslant l_2 - 2h_2, \\ g(x_1,\ l_2), & x_2 = l_2 - h_2. \end{cases}$$

Etudions les propriétés de l'opérateur A agissant de $H(\omega)$ vers $H(\omega)$.

1. L'opérateur A est autoadjoint:

$$(Au,\ v) = (u,\ Av), \quad u, v \in H(\omega). \tag{62}$$

Dans la démonstration tenons compte de ce que

$$(A_1 u,\ v) = (-\Lambda_1 \overset{\circ}{u},\ \overset{\circ}{v}) = -\sum_{j=1}^{N_2-1} h_2 \sum_{i=1}^{N_1-1} h_1 (\overset{\circ}{v}\Lambda_1 \overset{\circ}{u})_{ij} =$$

$$= -\sum_{j=1}^{N_2-1} h_2 \sum_{i=1}^{N_1-1} h_1 (\overset{\circ}{u}\Lambda_1 \overset{\circ}{v})_{ij} = -(\overset{\circ}{u},\ \Lambda_1 \overset{\circ}{v}) = (u,\ A_1 v),$$

car l'opérateur de différences Λ_1 en vertu de la seconde formule de différences de Green sur le maillage $\bar{\omega}_1 = \{x_1(i) = ih_1, 0 \leqslant i \leqslant N_1, h_1 N_1 = l_1\}$ satisfait à l'égalité

$$\sum_{i=1}^{N_1-1} h_1 (\overset{\circ}{v}\Lambda_1 \overset{\circ}{u})_{ij} = \sum_{i=1}^{N_1-1} h_1 (\overset{\circ}{u}_1 \Lambda_1 \overset{\circ}{v})_{ij},$$

en outre, il est possible de permuter l'ordre de sommation en i et j.

De façon analogue on obtient que $(A_2 u,\ v) = (u,\ A_2 v)$. Il s'ensuit (62).

2. L'opérateur A est défini positif et satisfait à l'estimation

$$\delta E \leqslant A \leqslant \Delta E, \quad \delta > 0, \tag{63}$$

où

$$\delta = \sum_{\alpha=1}^{2} \frac{4}{h_\alpha^2} \sin^2 \frac{\pi}{2N_\alpha} \geqslant \sum_{\alpha=1}^{2} \frac{8}{l_\alpha^2}, \quad \Delta = \sum_{\alpha=1}^{2} \frac{4}{h_\alpha^2} \cos^2 \frac{\pi}{2N_\alpha} \leqslant \sum_{\alpha=1}^{2} \frac{4}{h_\alpha^2}. \tag{64}$$

Notons que δ et Δ sont des valeurs propres minimale et maximale de l'opérateur de différences de Laplace Λ (voir point 1, § 2, ch. IV).

Cette assertion se démontre de la même façon que pour le lemme 12. On a donc établi que dans $H = H(\omega)$

$$A = A^*, \quad \delta E \leqslant A \leqslant \Delta E, \quad \delta > 0.$$

Si à la partie γ_0 de la maille frontière γ est imposée la condition aux limites de première espèce $y(x) = g(x)$, $x \in \gamma_0$, tandis qu'à la partie restante sont imposées les conditions aux limites de deuxième ou de troisième espèces, l'opérateur A se définit alors au moyen de la méthode décrite plus haut, $\overset{\circ}{H}$ étant l'ensemble des fonctions qui ne deviennent nulles que sur γ_0, tandis que $H = H(\omega_0)$ constitue l'espace des fonctions de mailles données sur $\omega_0 = \omega \cup (\gamma \setminus \gamma_0)$. Soit par exemple $\gamma_0 = \{x_{ij} \in \omega,\ i = 0,\ 0 \leqslant j \leqslant N_2\}$, tandis que sur $\gamma \setminus \gamma_0$ sont données les conditions aux limites de seconde espèce. Le schéma aux différences s'écrit alors sous forme

$$\Lambda y = (\Lambda_1 + \Lambda_2) y = -\varphi(x), \quad x \in \omega_0,$$
$$y(x) = g(x), \quad x \in \gamma_0.$$

Dans ce cas

$$\Lambda_2 y = \begin{cases} \dfrac{2}{h_2} y_{x_2}, & x_2 = 0, \\ y_{\bar{x}_2 x_2}, & h_2 \leqslant x_2 \leqslant l_2 - h_2, \\ -\dfrac{2}{h_2} y_{\bar{x}_2}, & x_2 = l_2, \quad h_1 \leqslant x_1 \leqslant l_1, \end{cases}$$

tandis que l'opérateur Λ_1 est défini par les formules

$$\Lambda_1 y = \begin{cases} y_{\bar{x}_1 x_1}, & h_1 \leqslant x_1 \leqslant l_1 - h_1, \\ -\dfrac{2}{h_1} y_{\bar{x}_1}, & x_1 = l_1, \quad 0 \leqslant x_2 \leqslant l_2. \end{cases}$$

Le produit scalaire dans l'espace $H = H(\omega_0)$ se définit par la formule

$$(u, v) = \sum_{i=1}^{N_1} \sum_{j=0}^{N_2} u(i, j)\, v(i, j)\, \hbar_1(i)\, \hbar_2(j),$$

où

$$\hbar_1(i) = \begin{cases} h_1, & 1 \leqslant i \leqslant N_1 - 1, \\ 0{,}5 h_1, & i = N_1, \end{cases}$$

$$\hbar_2(j) = \begin{cases} h_2, & 1 \leqslant j \leqslant N_2 - 1, \\ 0{,}5 h_2, & j = 0,\ N_2. \end{cases}$$

On peut montrer que l'opérateur $A = A_1 + A_2$ correspondant à l'opérateur de différences Λ est autoadjoint dans H et que pour ce dernier les estimations (63) avec $\delta = \delta_1 + \delta_2$, $\Delta = \Delta_1 + \Delta_2$, $\delta_1 = \dfrac{4}{h_1^2} \sin^2 \dfrac{\pi}{4N_1}$, $\Delta_1 = \dfrac{4}{h_1^2} \cos^2 \dfrac{\pi}{4N_1}$, $\delta_2 = 0$, $\Delta_2 = \dfrac{4}{h_2^2}$ sont vérifiées. δ_α et Δ_α sont ici les valeurs propres minimale et maximale de l'opérateur de différences Λ_α, $\alpha = 1, 2$.

Remarquons que les opérateurs A_1 et A_2 sont permutables aussi bien pour le premier que pour le second problème aux limites. Aussi

en vertu de la théorie générale (voir point 5, § 1, ch. V) les valeurs propres de l'opérateur A sont-elles la somme des valeurs propres des opérateurs A_1 et A_2: $\lambda(A) = \lambda(A_1) + \lambda(A_2)$.

8. Equations avec coefficients variables et avec dérivées mixtes. Examinons le problème de Dirichlet pour l'équation elliptique à coefficients variables dans le rectangle $\overline{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha,\ \alpha = 1, 2\}$:

$$Lu = \sum_{\alpha=1}^{2} \frac{\partial}{\partial x_\alpha}\left(k_\alpha(x)\frac{\partial u}{\partial x_\alpha}\right) - q(x)\,u = -\varphi(x), \quad x \in G,$$
$$u(x) = g(x), \quad x \in \Gamma, \tag{65}$$

où $k_\alpha(x)$ et $q(x)$ sont des fonctions suffisamment lisses satisfaisant aux conditions $0 < c_1 \leqslant k_\alpha(x) \leqslant c_2$, $0 \leqslant d_1 \leqslant q(x) \leqslant d_2$. Désignons par $\overline{\omega} = \omega + \gamma$ le maillage de pas h_1 et h_2 introduit au point 7.

Le problème (65) sera mis en accord avec le problème de différences de Dirichlet sur le maillage $\overline{\omega}$:

$$\Lambda y = (\Lambda_1 + \Lambda_2)\,y - dy = -\varphi(x), \quad x \in \omega,$$
$$y(x) = g(x), \quad x \in \gamma, \tag{66}$$

où $\Lambda_\alpha y = (a_\alpha y_{\bar{x}_\alpha})_{x_\alpha}$, $\alpha = 1, 2$, tandis que $a_\alpha(x)$ et $d(x)$ sont choisis, par exemple, ainsi:

$$a_1(x_1, x_2) = k_1(x_1 - 0.5h_1, x_2),$$
$$a_2(x_1, x_2) = k_2(x_1, x_2 - 0.5h_2), \quad d(x) = q(x).$$

Dans ce cas les coefficients du schéma aux différences remplissent les conditions

$$0 < c_1 \leqslant a_\alpha(x) \leqslant c_2, \quad 0 \leqslant d_1 \leqslant d \leqslant d_2. \tag{67}$$

Désignons par $H = H(\omega)$ l'espace des fonctions de mailles introduit au point précédent et par $\overset{\circ}{H}$ l'ensemble des fonctions de mailles s'annulant sur γ.

Ecrivons le schéma aux différences (66) sous forme d'équation opératorielle (59), où l'opérateur A est défini de façon triviale: $Ay = -\Lambda\overset{\circ}{y}$ avec $y \in H$, $\overset{\circ}{y} \in \overset{\circ}{H}$ et $y(x) = \overset{\circ}{y}(x)$ pour $x \in \omega$.

Désignons par $\mathcal{R} = \mathcal{R}_1 + \mathcal{R}_2$, où $\mathcal{R}_\alpha y = y_{\bar{x}_\alpha x_\alpha}$, $\alpha = 1, 2$, l'opérateur de différences de Laplace et définissons dans H l'opérateur R qui lui correspond: $Ry = -\mathcal{R}\overset{\circ}{y}$, $y \in H$, $\overset{\circ}{y} \in H$ et $y(x) = \overset{\circ}{y}(x)$ pour $x \in \omega$.

L e m m e 19. *L'opérateur A est autoadjoint dans H et satisfait aux estimations*

$$(c_1 + d_1/\Delta)\,(Ru,\ u) \leqslant (Au,\ u) \leqslant (c_2 + d_2/\delta)\,(Ru,\ u). \qquad (68)$$
$$(c_1\delta + d_1)\,(u,\ u) \leqslant (Au,\ u) \leqslant (c_2\Delta + d_2)\,(u,\ u), \qquad (69)$$

où δ et Δ sont définis dans (64).

En effet, à partir des conditions (67) et des estimations obtenues au point précédent

$$\delta E \leqslant R \leqslant \Delta E. \qquad (70)$$

il s'ensuit que pour tout $u \in H$ se vérifient les inégalités

$$\frac{d_1}{\Delta}(Ru,\ u) \leqslant d_1(u,\ u) \leqslant (du,\ u) \leqslant d_2(u,\ u) \leqslant \frac{d_2}{\delta}(Ru,\ u). \qquad (71)$$

Ensuite, la première formule de différences de Green donne

$$(A_1u,\ u) = -(\Lambda_1\mathring{u},\ \mathring{u}) = \sum_{j=1}^{N_2-1}\sum_{i=1}^{N_1}(a_1\mathring{u}^2_{\bar{x}_1})_{ij}\,h_1h_2.$$
$$(R_1u,\ u) = -(\mathcal{R}_1\mathring{u},\ \mathring{u}) = \sum_{j=1}^{N_2-1}\sum_{i=1}^{N_1}(\mathring{u}^2_{\bar{x}_1})_{ij}\,h_1h_2.$$

En vertu de (67) il s'ensuit l'inégalité

$$c_1\,(R_1u,\ u) \leqslant (A_1u,\ u) \leqslant c_2\,(R_1u,\ u).$$

De façon analogue on aboutit à

$$c_1\,(R_2u,\ u) \leqslant (A_2u,\ u) \leqslant c_2\,(R_2u,\ u).$$

De là, ainsi que de (70), on déduit les inégalités

$$c_1\delta\,(u,\ u) \leqslant c_1\,(Ru,\ u) \leqslant ((A_1 + A_2)\,u,\ u) \leqslant c_2\,(Ru,\ u) \leqslant c_2\Delta\,(u,\ u),$$

qui une fois additionnées avec les inégalités (71) donnent (68) et (69).

Le fait que l'opérateur A est autoadjoint se démontre par analogie avec le point précédent.

Notons que dans les inégalités (68) figurent les constantes de l'équivalence énergétique des opérateurs R et A, en outre, comme $d_1 \geqslant 0$ et $\delta \geqslant 8/l_1^2 + 8/l_2^2$ ces opérateurs sont équivalents aux constantes qui sont indépendantes du nombre de nœuds dans le maillage.

Examinons maintenant le *problème de Dirichlet pour l'équation elliptique renfermant des dérivées mixtes*

$$Lu = \sum_{\alpha,\,\beta=1}^{2}\frac{\partial}{\partial x_\alpha}\left(k_{\alpha\beta}(x)\,\frac{\partial u}{\partial x_\beta}\right) = -\varphi(x), \quad x \in \overline{G},$$
$$u(x) = g(x). \qquad\qquad x \in \Gamma. \qquad (72)$$

On admet par hypothèse que les conditions d'ellipticité sont remplies

$$c_1 \sum_{\alpha=1}^{2} \xi_\alpha^2 \leqslant \sum_{\alpha,\beta=1}^{2} k_{\alpha\beta}(x)\,\xi_\alpha\xi_\beta \leqslant c_2 \sum_{\alpha=1}^{2} \xi_\alpha^2, \quad x \in \overline{G}, \tag{73}$$

où $c_2 \geqslant c_1 > 0$, tandis que $\xi = (\xi_1, \xi_2)$ est un vecteur quelconque.

Sur un maillage rectangulaire $\overline{\omega}$ on peut opposer au problème (72) le schéma aux différences

$$\Lambda y = 0{,}5 \sum_{\alpha,\beta=1}^{2} [(k_{\alpha\beta} y_{\bar{x}_\beta})_{x_\alpha} + (k_{\alpha\beta} y_{x_\beta})_{\bar{x}_\alpha}] = -\varphi(x), \quad x \in \omega,$$
$$y(x) = g(x), \quad x \in \gamma. \tag{74}$$

Ecrivons (74) sous forme de l'équation opératorielle (59) en définissant de façon triviale l'opérateur A: $Ay = -\Lambda \overset{\circ}{y}$, où $y \in H(\omega)$, $\overset{\circ}{y} \in \overset{\circ}{H}$ et $y(x) = \overset{\circ}{y}(x)$ pour $x \in \omega$. De plus, le second membre f ne diffère du second membre φ de l'équation (74) qu'aux nœuds frontières. Pour expliciter f, il faut transcrire l'équation aux différences dans le nœud frontière, utiliser les conditions aux limites et rapporter dans le second membre de l'équation les valeurs connues de $y(x)$ sur γ.

Montrons maintenant qu'avec la réalisation des conditions de symétrie $k_{12}(x) = k_{21}(x)$ l'opérateur A devient autoadjoint dans l'espace $H = H(\omega)$ défini plus haut. A cette fin écrivons l'opérateur Λ sous forme de somme $\Lambda = (\Lambda_1 + \Lambda_2)/2$, où

$$\Lambda_\alpha y = (k_{\alpha\alpha} y_{\bar{x}_\alpha} + k_{\alpha\beta} y_{\bar{x}_\beta})_{x_\alpha} + (k_{\alpha\alpha} y_{x_\alpha} + k_{\alpha\beta} y_{x_\beta})_{\bar{x}_\alpha},$$
$$\beta = 3 - \alpha, \quad \alpha = 1, 2.$$

En utilisant les formules de sommation par parties (7′) et (9′), on obtient pour tous $\overset{\circ}{u}, \overset{\circ}{v} \in \overset{\circ}{H}$

$$(\Lambda_1 \overset{\circ}{u}, \overset{\circ}{v}) = -\sum_{j=1}^{N_2-1} \sum_{i=1}^{N_1} [(k_{11} \overset{\circ}{u}_{\bar{x}_1} + k_{12} \overset{\circ}{u}_{\bar{x}_2})\, \overset{\circ}{v}_{\bar{x}_1}]_{ij} h_1 h_2 -$$
$$- \sum_{j=1}^{N_2-1} \sum_{i=0}^{N_1-1} [(k_{11} \overset{\circ}{u}_{x_1} + k_{12} \overset{\circ}{u}_{x_2})\, \overset{\circ}{v}_{x_1}]_{ij}\, h_1 h_2.$$

En tenant compte de ce que $\overset{\circ}{v}_{\bar{x}_1}$ et $\overset{\circ}{v}_{x_1}$ sont nuls pour $j = N_2$ et $j = 0$, l'égalité obtenue peut être écrite sous la forme

$$(\Lambda_1 \overset{\circ}{u}, \overset{\circ}{v}) = -\sum_{j=1}^{N_2} \sum_{i=1}^{N_1} [(k_{11} \overset{\circ}{u}_{\bar{x}_1} + k_{12} \overset{\circ}{u}_{\bar{x}_2})\, \overset{\circ}{v}_{\bar{x}_1}]_{ij}\, h_1 h_2 -$$
$$- \sum_{j=0}^{N_2-1} \sum_{i=0}^{N_1-1} [(k_{11} \overset{\circ}{u}_{x_1} + k_{12} \overset{\circ}{u}_{x_2})\, \overset{\circ}{v}_{x_1}]_{ij}\, h_1 h_2. \tag{75}$$

De façon analogue, on obtient

$$(\Lambda_2\overset{\circ}{u}, \overset{\circ}{v}) = -\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}[(k_{22}\overset{\circ}{u}_{\bar{x}_2} + k_{21}\overset{\circ}{u}_{\bar{x}_1})\overset{\circ}{v}_{\bar{x}_2}]_{ij}h_1h_2 -$$

$$-\sum_{i=0}^{N_1-1}\sum_{j=0}^{N_2-1}[(k_{22}\overset{\circ}{u}_{x_2} + k_{21}\overset{\circ}{u}_{x_1})\overset{\circ}{v}_{x_2}]_{ij}h_1h_2. \quad (76)$$

En additionnant (75) et (76), il vient

$$(\Lambda\overset{\circ}{u}, \overset{\circ}{v}) = -0{,}5\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}h_1h_2\Big(\sum_{\alpha,\beta=1}^{2}k_{\alpha\beta}\overset{\circ}{u}_{\bar{x}_\alpha}\overset{\circ}{v}_{\bar{x}_\beta}\Big)_{ij} -$$

$$-0{,}5\sum_{i=0}^{N_1-1}\sum_{j=0}^{N_2-1}h_1h_2\Big(\sum_{\alpha,\beta=1}^{2}k_{\alpha\beta}\overset{\circ}{u}_{x_\alpha}\overset{\circ}{v}_{x_\beta}\Big)_{ij}. \quad (77)$$

Il s'ensuit que si $k_{12} = k_{21}$, on a l'égalité

$$(\Lambda\overset{\circ}{u}, \overset{\circ}{v}) = (\overset{\circ}{u}, \Lambda\overset{\circ}{v}).$$

En vertu de l'égalité $(Au, v) = -(\Lambda\overset{\circ}{u}, \overset{\circ}{v})$ *l'opérateur A est autoadjoint dans H.*

Cherchons les bornes de l'opérateur A. En portant dans (77) au lieu de $\overset{\circ}{v}$ la fonction de maille $\overset{\circ}{u}$, tenons compte de l'ellipticité (73) et de la condition $\overset{\circ}{u}(x) = 0$ pour $x \in \gamma$. Il vient

$$-(\Lambda\overset{\circ}{u}, \overset{\circ}{u}) \geqslant 0{,}5c_1\Big\{\sum_{j=1}^{N_2-1}h_2\Big[\sum_{i=1}^{N_1}(\overset{\circ}{u}_{\bar{x}_1})_{ij}^2h_1 + \sum_{i=0}^{N_1-1}(\overset{\circ}{u}_{x_1})_{ij}^2h_1\Big] +$$

$$+\sum_{i=1}^{N_1-1}h_1\Big[\sum_{j=1}^{N_2}(\overset{\circ}{u}_{\bar{x}_2})_{ij}^2h_2 + \sum_{j=0}^{N_2-1}(\overset{\circ}{u}_{x_2})_{ij}^2h_2\Big]\Big\} =$$

$$= c_1\Big[\sum_{j=1}^{N_2-1}\sum_{i=1}^{N_1}(\overset{\circ}{u}_{\bar{x}_1})_{ij}^2h_1h_2 + \sum_{i=1}^{N_1-1}\sum_{j=1}^{N_2}(\overset{\circ}{u}_{\bar{x}_2})_{ij}^2h_1h_2\Big] = c_1(-\mathscr{H}\overset{\circ}{u}, \overset{\circ}{u}),$$

où $\mathscr{H}$ est l'opérateur de différences de Laplace. De façon analogue on obtient

$$-(\Lambda\overset{\circ}{u}, \overset{\circ}{u}) \leqslant c_2(-\mathscr{H}\overset{\circ}{u}, \overset{\circ}{u}).$$

Compte tenu de l'estimation (70), on aboutit aux inégalités suivantes relativement à l'opérateur A :

$$\begin{gathered} c_1(Ru, u) \leqslant (Au, u) \leqslant c_2(Ru, u), \\ c_1\delta(u, u) \leqslant (Au, u) \leqslant c_2\Delta(u, u), \end{gathered} \quad (78)$$

où δ et Δ sont définis dans (64). Par conséquent, l'opérateur A correspondant à l'opérateur de différences elliptique avec dérivées mixtes et l'opérateur R correspondant à l'opérateur de différences de Laplace sont énergiquement équivalents aux constantes c_1 et c_2 indépendantes du nombre de nœuds dans le maillage. L'opérateur A

possède des bornes $c_1\delta = O(1)$ et $c_2\Delta = O(1/h^2)$ $(h^2 = h_1^2 + h_2^2)$ et si le nombre de nœuds du maillage est grand l'opérateur A est mal conditionné.

Notons que les inégalités (78) restent vraies même au cas où pour l'approximation de l'opérateur différentiel L on utilise des opérateurs de différences

$$\Lambda y = \frac{1}{2} \sum_{\alpha=1}^{2} [(k_{\alpha\alpha} y_{\bar{x}_\alpha})_{x_\alpha} + (k_{\alpha\alpha} y_{x_\alpha})_{\bar{x}_\alpha}] + \frac{1}{2} \sum_{\alpha \neq \beta}^{1\div 2} [(k_{\alpha\beta} y_{x_\beta})_{x_\alpha} + (k_{\alpha\beta} y_{\bar{x}_\beta})_{\bar{x}_\alpha}]$$

ou

$$\Lambda y = \frac{1}{2} \sum_{\alpha=1}^{2} [(k_{\alpha\alpha} y_{\bar{x}_\alpha})_{x_\alpha} + (k_{\alpha\alpha} y_{x_\alpha})_{\bar{x}_\alpha}] + \frac{1}{4} \sum_{\alpha \neq \beta}^{1\div 2} [(k_{\alpha\beta} y_{\bar{x}_\beta})_{x_\alpha} + (k_{\alpha\beta} y_{x_\beta})_{\bar{x}_\alpha} + (k_{\alpha\beta} y_{x_\beta})_{x_\alpha} + (k_{\alpha\beta} y_{\bar{x}_\beta})_{\bar{x}_\alpha}].$$

§ 3. Notions générales sur la théorie des méthodes itératives

1. Méthode de stationnarisation. On a montré plus haut que les schémas aux différences des équations elliptiques se transcrivent de façon naturelle sous forme d'équation opératorielle de première espèce

$$Au = f \tag{1}$$

dont l'opérateur A agit dans l'espace hilbertien H de dimension finie. Aux équations elliptiques linéaires correspondent des opérateurs A linéaires, et aux équations quasi linéaires des opérateurs A non linéaires.

La théorie des méthodes itératives de l'équation opératorielle (1) peut être exposée comme une des branches de la théorie générale de stabilité des schémas aux différences. Les schémas itératifs peuvent être assimilés à des méthodes de stationnarisation de l'équation non stationnaire correspondante. Eclairons-le sur un exemple d'équation à opérateur A autoadjoint, défini positif et borné, $A = A^* \geqslant \delta E$, $\delta > 0$.

Soit $v = v(t)$ une fonction abstraite de t à valeurs dans H, c'est-à-dire que $v(t)$ est un élément de l'espace H pour chaque t fixé. Etudions le problème abstrait de Cauchy:

$$\frac{dv}{dt} + Av = f, \quad t > 0, \quad v(0) = v_0 \in H. \tag{2}$$

Montrons que $\lim\limits_{t\to\infty} \| v(t) - u \| = 0$, où u est la solution de l'équation (1), autrement dit la solution $v(t)$ de l'équation non stationnaire (2) tend avec l'accroissement de t vers la solution u de l'équation stationnaire (indépendante de t) (1) (il y a lieu à « stationnarisation » ou à une « sortie sur un régime stationnaire »). Pour une erreur $z(t) = v(t) - u$ on a une équation homogène

$$\frac{dz}{dt} + Az = 0, \quad t > 0, \quad z(0) = v(0) - u.$$

En multipliant cette équation scalairement par z: $\left(\frac{dz}{dt}, z\right) + (Az, z) = 0$ et compte tenu de

$$\left(\frac{dz}{dt}, z\right) = \frac{1}{2}\frac{d}{dt}(z, z) = \frac{1}{2}\frac{d}{dt}\| z \|^2, \quad (Az, z) \geqslant \delta \| z \|^2,$$

il vient

$$\frac{d}{dt}\| z(t) \|^2 + 2\delta \| z(t) \|^2 \leqslant 0.$$

Après multiplication de cette inégalité par $e^{2\delta t} > 0$, on a

$$\frac{d}{dt} e^{2\delta t} \| z(t) \|^2 \leqslant 0,$$

d'où il s'ensuit que $e^{2\delta t} \| z(t) \|^2 \leqslant \| z(0) \|^2$ ou

$$\| v(t) - u \| \leqslant e^{-\delta t} \| v(0) - u \| \to 0 \quad \text{pour } t \to \infty.$$

Donc en résolvant l'équation (2) pour tout $v_0 \in H$, on obtient, au cas de t suffisamment grand, la solution approchée de l'équation initiale (1) à toute précision voulue. Ce procédé d'obtention de la solution est appelé *méthode de stationnarisation*. Une propriété analogue d'amortissement des données initiales est propre aux analogues de l'équation (2) au sens des différences finies.

2. Schémas itératifs. Arrêtons-nous d'abord sur la caractéristique générale de la notion de schéma itératif. Supposons qu'il s'agit de trouver la solution de l'équation (1). Admettons tout d'abord que A est un opérateur linéaire défini dans H.

Dans toute méthode itérative de résolution de l'équation (1) on part d'une certaine approximation initiale $y_0 \in H$ en déterminant de proche en proche les solutions approchées $y_1, y_2, \ldots, y_k, y_{k+1}, \ldots$, où k est le numéro de l'itération. L'approximation y_{k+1} est exprimée en fonction des approximations déjà connues au moyen de la formule de récurrence

$$y_{k+1} = F_k(y_0, y_1, \ldots, y_k),$$

où F_k est une certaine fonction dépendant en général de l'opérateur A, du second membre f, du numéro d'itération k.

On dit que la méthode intérative est de l'ordre m si chaque approximation suivante ne dépend que des m approximations précédentes. c'est-à-dire

$$y_{k+1} = F_k (y_{k-m+1}, y_{k-m+2}, \ldots, y_k).$$

Les schémas itératifs d'ordre élevé exigent pour leur réalisation la mémorisation d'un énorme volume d'information intermédiaire, aussi en pratique se limite-t-on à des valeurs de $m = 1$ ou $m = 2$.

Du choix de la fonction F_k dépend la structure du schéma itératif. Si la fonction est linéaire, la méthode itérative est également dite linéaire. Si F_k est indépendant du numéro d'itération k, la méthode itérative est dite stationnaire.

Etudions l'aspect général du schéma itératif linéaire du premier ordre. Tout schéma de ce genre, en accord avec la définition, peut être écrit sous la forme

$$y_{k+1} = S_{k+1} y_k + \tau_{k+1} \varphi_{k+1}, \quad k = 0, 1, \ldots, \tag{3}$$

où S_k est l'opérateur linéaire donné sur H, τ_k certains paramètres numériques.

Généralement on exige des schémas itératifs une condition toute naturelle: la solution $u = A^{-1}f \in H$ de l'équation (1) doit être pour tout f un point immobile du procédé d'approximations successives (3), autrement dit

$$A^{-1}f = S_{k+1} A^{-1}f + \tau_{k+1} \varphi_{k+1}. \tag{4}$$

Il s'ensuit que si l'on pose

$$S_{k+1} = E - \tau_{k+1} B_{k+1}^{-1} A, \quad \varphi_{k+1} = B_{k+1}^{-1} f, \tag{5}$$

où B_{k+1} est un opérateur linéaire inversible agissant dans H, la condition (4) sera remplie. En portant (5) dans (3), on obtient finalement après quelques transformations fort simples

$$B_{k+1} \frac{y_{k+1} - y_k}{\tau_{k+1}} + A y_k = f, \quad k = 0, 1, \ldots, \quad y_0 \in H. \tag{6}$$

En se conformant à la terminologie de la théorie des schémas aux différences [voir A. Samarski, Théorie des schémas aux différences. 1977, ch. V (en russe)], appelons (6) forme canonique du schéma itératif à deux couches. Bref, tout processus itératif linéaire du premier ordre peut être transcrit sous la forme (6). Si $B_{k+1} \equiv E$, le schéma itératif est appelé *explicite*, vu que dans ce cas l'approximation y_{k+1} est de forme explicite

$$y_{k+1} = y_k - \tau_{k+1} (A y_k - f), \quad k = 0, 1, \ldots$$

Si B_k diffère au moins pour un k de l'opérateur unitaire, le schéma est dit *implicite*. Les nombres τ_k sont appelés paramètres d'itération. Si τ_{k+1} dépend de l'approximation itérative y_k, *le processus itératif sera non linéaire*. Il est évident que dans le processus itératif station-

naire les opérateurs B_k et les paramètres τ_k (plus précisément B_k/τ_{k+1}) ne doivent pas dépendre du numéro d'itération k.

Notons que le schéma (6) peut être traité comme un schéma implicite à deux couches de l'équation non stationnaire

$$B(t)\frac{dv}{dt}+Av=f, \quad t>0, \; v(0)=y_0,$$

de nature plus générale que l'équation (2) étudiée plus haut. De plus, le paramètre τ_{k+1} peut être assimilé au pas par rapport au temps fictif.

La différence entre les schémas itératifs et les schémas pour problèmes non stationnaires de la forme (2) est la suivante:

1) pour tous B_{k+1} et τ_{k+1} la solution u de l'équation initiale (1) satisfait à (6);

2) le choix des paramètres τ_{k+1} et des opérateurs B_{k+1} ne doit se plier qu'aux exigences de convergence des itérations et du minimum d'opérations arithmétiques que coûte la recherche de la solution de l'équation (1) avec une précision donnée (pour les problèmes non stationnaires le choix du pas est avant tout assujetti à la nécessité d'approximation).

On a admis plus haut que l'opérateur A était linéaire. Le schéma (6) peut, apparemment, être utilisé à la recherche de la solution approchée de (1) également au cas où l'opérateur A est non linéaire. Pour ce faire, l'opérateur B_{k+1} est habituellement choisi linéaire.

Les schémas itératifs à deux couches sont les plus utilisés. Cependant dans la résolution de l'équation (1) on utilise également des schémas à trois couches qui décrivent les processus itératifs du second ordre. Les schémas à trois couches les plus étudiés sont les schémas du type « standard ». Ils se transcrivent sous forme

$$B_{k+1}y_{k+1}=\alpha_{k+1}(B_{k+1}-\tau_{k+1}A)y_k+ \\ +(1-\alpha_{k+1})B_{k+1}y_{k-1}+\alpha_{k+1}\tau_{k+1}f \qquad (7)$$

pour $k=1, 2, \ldots$ On utilise ici deux suites de paramètres itératifs $\{\tau_k\}$ et $\{\alpha_k\}$. Pour la mise en œuvre du schéma (7) il faut, outre l'approximation initiale y_0, définir encore l'approximation y_1. Habituellement, on l'obtient à partir de y_0 en utilisant le schéma à deux couches (6), c'est-à-dire

$$B_1y_1=(B_1-\tau_1A)y_0+\tau_1f, \quad y_0\in H. \qquad (8)$$

On peut montrer que pour (7), (8) la solution u de l'équation (1) est un point immobile.

Si $B_k\equiv E$ pour tous $k=1, 2, \ldots$, le schéma (7) est alors dit *explicite*:

$$y_{k+1}=\alpha_{k+1}(E-\tau_{k+1}A)y_k+(1-\alpha_{k+1})y_{k-1}+\alpha_{k+1}\tau_{k+1}f.$$

Dans le cas contraire le schéma (7) est *implicite*.

8. Convergence et nombre d'itérations. La principale différence entre les méthodes itératives et directes réside dans le fait que les méthodes itératives ne fournissent une solution précise de l'équation (1) que comme la limite d'une suite d'approximations itératives $\{y_k\}$ pour $k \to \infty$. Font exception les méthodes d'itérations « finies » auxquelles se rapportent les méthodes de directions conjuguées qui, théoriquement, permettent d'obtenir la solution précise pour toute approximation initiale en un nombre d'opérations fini, si A est un opérateur linéaire dans un espace de dimensions finies.

Pour caractériser l'écart de l'approximation itérative y_k de la solution précise u du problème (1), on introduit l'*erreur* $z_k = y_k - u$. Le processus itératif est dit *convergent dans l'espace énergétique* H_D, si $\| z_k \|_D \to 0$ pour $k \to \infty$. H_D est ici l'espace engendré par l'opérateur D autoadjoint et défini positif dans H.

La raison de l'introduction de l'espace énergétique H_D est la suivante. Comme on le sait, la suite des éléments H convergeant dans une norme converge également dans une norme équivalente. Aussi dans l'étude d'un schéma itératif concret est-il commode de choisir un tel espace énergétique H_D dans lequel les opérateurs du schéma itératif A et B_k soient munis de propriétés imposées, par exemple, seraient autoadjoints et définis positifs.

Une des caractéristiques essentielles de la méthode itérative est le nombre d'itérations. Habituellement on fixe une certaine précision $\varepsilon > 0$ avec laquelle il s'agit de trouver la solution approchée de l'équation (1). Si $\| u \|_D = O(1)$, il faut que soit remplie la condition

$$\| y_n - u \|_D \leqslant \varepsilon \quad \text{pour} \quad n \geqslant n_0(\varepsilon). \tag{9}$$

$n_0(\varepsilon)$ est le nombre minimal d'itérations garantissant la précision donnée ε. Ce nombre est fonction du choix de l'approximation initiale. La condition (9) permet de déterminer le moment de la fin des itérations, au cas où la norme indiquée se prête efficacement au calcul au cours des itérations. Par exemple, si l'opérateur A est non dégénéré et défini positif, en choisissant pour D l'opérateur A^*A, on obtient de (9)

$$\| y_n - u \|_D = \| Ay_n - f \| \leqslant \varepsilon,$$

car

$$(y_n - u,\ y_n - u)_D = (A^*A\ (y_n - u),\ y_n - u) =$$
$$= (Ay_n - Au,\ Ay_n - Au) = \| Ay_n - f \|^2.$$

Pour la comparaison de la qualité des différentes méthodes, on se réfère généralement au nombre d'itérations qu'on déduit de la condition

$$\| y_n - u \|_D \leqslant \varepsilon \| y_0 - u \|_D \quad \text{pour} \quad n \geqslant n_0(\varepsilon). \tag{10}$$

Ce nombre indique le nombre d'itérations qu'il suffit de réaliser pour que pour toute approximation initiale y_0 la norme de l'erreur

initiale dans H_D soit réduite de $1/\varepsilon$ fois. La condition (10) peut également être utilisée en guise de critère d'achèvement du processus d'itérations.

On peut opposer à l'équation (1) un grand nombre de schémas itératifs (6) ou (7), (8) avec B_k et τ_k, α_k quelconques. Toutefois lors de la résolution d'un problème concret on voit apparaître le problème du choix d'un schéma unique. Sous l'angle du calcul mathématique, l'essentiel est de construire des méthodes itératives capables d'aboutir à la solution de (1) avec la précision voulue en un temps machine minimal. Cette exigence envers la rentabilité de la méthode est toute naturelle. Lors des appréciations théoriques de la qualité de la méthode, elle est souvent remplacée par le critère du minimum d'opérations arithmétiques $Q(\varepsilon)$ permettant d'obtenir la solution avec la précision voulue.

Le volume total de calcul $Q(\varepsilon)$ vaut $Q(\varepsilon) = \sum_{k=1}^{n} q_k$, où q_k est le nombre d'opérations de calcul de l'itération de numéro k, tandis que n est le nombre d'itérations, $n \geqslant n_0(\varepsilon)$. Le problème de construction de la méthode itérative se pose ainsi (pour un schéma à deux couches (6)) : l'opérateur A est fixé, tandis que les paramètres $\{\tau_k, k = 1, 2, \ldots, n\}$ et les opérateurs B_k doivent être choisis sur la base de la condition du minimum $Q(\varepsilon)$.

Ainsi posé, le problème n'a apparemment pas de solution. Habituellement la composition des opérateurs B_k est donnée à priori et si le nombre d'opérations nécessaire à l'inversion de l'opérateur B_k est indépendant de k, on a alors $q_k \equiv q$ et $Q(\varepsilon) = qn_0(\varepsilon)$. Dans ce cas le problème du minimum $Q(\varepsilon)$ se réduit au problème du choix des paramètres d'itération τ_k à partir de la condition du minimum du nombre d'itérations $n_0(\varepsilon)$.

Pour établir une hiérarchie des méthodes, il est nécessaire de les classer suivant une caractéristique quelconque. On recourt quelquefois à des estimations asymptotiques du nombre d'opérations ou du nombre d'itérations quand le nombre d'inconnues tend dans le schéma aux différences vers l'infini. Or, en fait, il existe une limite du nombre d'inconnues lors de la résolution des équations elliptiques à plusieurs dimensions par la méthode des différences finies. C'est ainsi que pour l'équation tridimensionnelle de Poisson le nombre moyen de nœuds pour chaque variable $N \approx 100$ nous place en face d'un système d'équations algébriques linéaires à $M = 10^6$ inconnues. Il semble peu logique d'augmenter le nombre de nœuds. Aussi la comparaison des méthodes doit-elle avant tout s'effectuer avec des schémas réels.

4. Classification des méthodes itératives. Les méthodes itératives se caractérisent par la structure des schémas itératifs, l'espace énergétique H_D dans lequel est étudiée la convergence de la méthode,

le type de la méthode itérative, la condition de l'achèvement du processus d'itérations, ainsi que par l'algorithme de la mise en œuvre d'un pas itératif.

On n'étudiera que les schémas itératifs à deux et à trois couches. explicites et implicites, pour lesquels la condition de l'achèvement du processus d'itérations sera la condition

$$\| y_n - u \|_D \leqslant \varepsilon \, \| y_0 - u \|_D, \quad \varepsilon > 0.$$

Dans la théorie générale des méthodes itératives on étudie deux types de méthodes: celles utilisant une information à priori sur les opérateurs du schéma itératif et celles qui ne l'utilisent pas (méthodes du type variationnel). Dans le premier cas les paramètres d'itération τ_k pour le schéma (6) et τ_k, α_k pour le schéma (7), (8) sont choisis sur la base de la condition du minimum, soit à partir de la norme de l'opérateur résolvant (opérateur reliant les approximations initiale et finale), soit à partir de la norme de l'opérateur de passage d'une itération à l'autre. Les paramètres d'itération sont dans ce cas choisis de façon à assurer une vitesse maximale à la convergence pour la pire des approximations initiales. Dans les méthodes de ce type la qualité de l'approximation initiale n'est pas prise en compte.

Dans les méthodes du type variationnel les paramètres d'itération sont choisis sur la base de la condition du minimum de certaines fonctionnelles reliées à l'équation de départ. On choisit, par exemple, en guise de fonctionnelle la norme énergétique de l'erreur de la k-ème itération. Dans ce cas les paramètres d'itération dépendent des approximations itératives précédentes et possèdent la faculté de tenir compte de la qualité de l'approximation initiale.

Dans la théorie générale des méthodes itératives on s'abstient d'étudier la structure concrète des opérateurs du schéma itératif (on ne se sert en théorie que du minimum d'information sur les opérateurs, de nature fonctionnelle générale). Cela permet d'aboutir au but principal: formuler les principes généraux de construction des méthodes itératives optimales suivant la nature et la forme de l'information à priori sur le problème, ainsi que des exigences imposées au mode de résolution de ce problème. Ces exigences supplémentaires peuvent, par exemple, consister dans l'obligation de construire une méthode optimale non pas pour un problème, mais pour une série de problèmes possédant un même opérateur A et des seconds membres différents.

La prise en compte de la structure de l'opérateur du problème résolu permet, apparemment, de bâtir des méthodes itératives spéciales possédant des vitesses de convergence supérieures à celles des méthodes de la théorie générale. On y aboutit par un choix approprié des opérateurs B_k et des paramètres d'itération. Les méthodes spéciales ont un domaine d'application restreint.

Arrêtons-nous maintenant sur le rôle joué par les opérateurs B_k. Pour les schémas itératifs implicites le choix des opérateurs B_k doit être soumis à deux exigences: la garantie de convergence la plus rapide de la méthode et celle de simplicité et d'économie de l'inversion de ces opérateurs. Ces exigences sont contradictoires. En effet, si dans le schéma (6) on pose $B_1 = A$ et $\tau_1 = 1$, alors pour toute approximation initiale la solution de l'équation (1) peut être obtenue avec une seule itération. La vitesse de convergence dans ce cas est maximale, toutefois l'inversion d'un tel opérateur B_1 équivaut à la résolution du problème primitif.

Il s'avère, comme il le sera montré plus loin, qu'il n'est pas nécessaire de choisir l'opérateur B_k égal à l'opérateur A. Il suffit que les énergies de ces opérateurs soient proches. Cette exigence ouvre des perspectives de choix dans la classe des opérateurs B, dont l'énergie est proche de celle de l'opérateur A, ceux qui se prêtent à une facile inversion.

Actuellement, lors de la construction des méthodes itératives implicites, on recourt le plus souvent à l'approche suivante. L'opérateur B_{k+1} est donné de façon constructive sous forme explicite, ou bien l'approximation itérative y_{k+1} s'obtient par quelques calculs auxiliaires qui peuvent être interprétés comme une inversion implicite de l'opérateur B_{k+1}.

Dans le premier cas l'opérateur B_{k+1} est habituellement choisi sous forme de produit d'un certain nombre d'opérateurs facilement inversibles, de manière que l'opérateur B_{k+1} soit à certains égards proche de l'opérateur A. En outre, les opérateurs compris dans le produit peuvent de leur côté, dépendre des paramètres assimilés à des paramètres d'itération auxiliaires. Par exemple, si $B_k = (E + \omega_k A_1)(E + \omega_k A_2)$, où A_α sont des opérateurs, ω_k sont alors des nombres représentant les paramètres. Dans ce cas la variabilité de l'opérateur B_k ne se manifeste que dans la dépendance des paramètres mentionnés ω_k du numéro d'itération k. Avec une telle construction de l'opérateur B_k on garantit l'unicité du processus de calcul permettant de trouver la solution approchée à chaque itération.

Arrêtons-nous sur deux algorithmes permettant d'obtenir la nouvelle approximation y_{k+1} au cas où l'opérateur B_{k+1} est de forme factorisée. Soient $B_{k+1} = B^1_{k+1} B^2_{k+1} \ldots B^p_{k+1}$ et y_{k+1} obtenus suivant le schéma itératif à deux couches (6). Dans le premier algorithme on résout la suite des équations

$$B^1_{k+1} v^1 = F_{k+1}, \quad B^\alpha_{k+1} v^\alpha = v^{\alpha-1}, \qquad \alpha = 2, 3, \ldots, p, \qquad (11)$$

où $F_{k+1} = B_{k+1} y_k - \tau_{k+1}(Ay_k - f)$. On voit que $y_{k+1} = v^p$. Chacune des équations (11) se résout sans peine. L'algorithme n'exige pas la mémorisation de l'information intermédiaire qui une fois obtenue est aussitôt utilisée. Le défaut de l'algorithme est la néces-

sité de calcul de l'élément $B_{k+1}y_k$, procédure devenant parfois très laborieuse.

Le second algorithme a la forme du schéma avec correction :

$$y_{k+1} = y_k - \tau_{k+1}v^p, \tag{12}$$
$$B^1_{k+1}v^1 = Ay_k - f, \ B^{\alpha}_{k+1}v^{\alpha} = v^{\alpha-1}, \ \alpha = 2, 3, \ldots, p.$$

Dans ce cas il faut mémoriser en outre l'approximation itérative précédente y_k et la stocker jusqu'à l'obtention de la correction v^p.

Dans le second procédé de construction de la méthode itérative implicite on part, par exemple, du schéma de la correction (12) en cherchant la correction v^p sous forme de solution approchée de l'équation auxiliaire

$$R_{k+1}v = r_k, \quad r_k = Ay_k - f. \tag{13}$$

Posons que (13) se résout par un schéma itératif à deux couches quelconque. Alors l'erreur $z^m = v^m - v$ vérifie l'équation homogène

$$z^{m+1} = S_{m+1}z^m, \quad m = 0, 1, \ldots, p-1, \quad z^0 = v^0 - v,$$

où S_{m+1} est l'opérateur de passage de la m-ième à la $(m+1)$-ième itération. De là il vient

$$z^p = v^p - v = S_pS_{p-1} \ldots S_1z^0 = T_p(v^0 - v), \ T_p = \prod_{m=1}^{p} S_m,$$

où T_p est l'opérateur résolvant. En y portant $v = R^{-1}_{k+1}r_k$ et en posant $v^0 = 0$, on obtient

$$v^p = (E - T_p)\,R^{-1}_{k+1}r_k \quad \text{ou} \quad v^p = B^{-1}_{k+1}r_k, \tag{14}$$

où, au moyen de B_{k+1}, est désigné l'opérateur $R_{k+1}(E - T_p)^{-1}$.

Portons (14) dans (12) et l'on trouve que y_{k+1} vérifie le schéma à deux couches (6) muni de l'opérateur mentionné B_{k+1}. Si la norme de l'opérateur T_p est petite, l'opérateur B_{k+1} est « proche » de l'opérateur R_{k+1}. Aussi en guise de l'opérateur R_{k+1} est-il naturel de choisir un opérateur proche de A.

CHAPITRE VI

MÉTHODES ITÉRATIVES À DEUX COUCHES

On étudie dans ce chapitre les méthodes itératives à deux couches susceptibles de résoudre l'équation opératorielle $Au = f$. Les paramètres d'itération sont choisis sur la base d'une information à priori relative aux opérateurs du schéma itératif. Dans le § 1 on montre comment se pose le problème du choix des paramètres d'un schéma à deux couches. Dans les §§ 2 et 3 le problème est résolu pour le cas d'opérateurs autoadjoints. On recourt à la méthode de Tchébychev et à la méthode itérative simple. Le § 4 étudie quelques procédés de choix du paramètre d'itération au cas d'opérateurs non autoadjoints et suivant le volume de l'information à priori. Au § 5 sont donnés quelques exemples d'applications des méthodes construites à la résolution des équations de mailles.

§ 1. Position du problème sur le choix des paramètres d'itération

1. Famille de base des schémas itératifs. Au chapitre V on a montré que les problèmes de différences aux limites pour équations elliptiques constituent des systèmes spéciaux d'équations algébriques qui peuvent être assimilés à des équations opératorielles de première espèce

$$Au = f \tag{1}$$

dans l'espace hilbertien réel H. Dans quelques cas particuliers ces systèmes peuvent être résolus de façon efficace par des méthodes directes étudiées dans les chapitres I-IV. Dans le cas général l'une des méthodes approchées de résolution des équations de mailles elliptiques est la méthode itérative. On commencera l'étude des méthodes itératives par les méthodes à deux couches les plus simples, à savoir par la *méthode de Tchébychev* et la *méthode itérative simple*.

Pour la résolution approchée de l'équation (1) à opérateur linéaire non dégénéré A donné dans H, examinons le *schéma itératif implicite à deux couches*

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}}+Ay_k=f, \quad k=0,\ 1,\ \ldots, \tag{2}$$

avec approximation initiale quelconque $y_0 \in H$. $\{\tau_k\}$ est ici la suite des paramètres d'itération, quant à B, c'est un opérateur linéaire

non dégénéré quelconque agissant dans H. La question du meilleur choix de l'opérateur B sera étudiée séparément. Ici, on ne notera que l'opérateur B doit s'inverser facilement.

La convergence du schéma itératif (2) sera étudiée dans l'espace énergétique H_D engendré par l'opérateur D autoadjoint et arbitraire, défini positif dans H.

Comme l'opérateur B n'est pas fixé, (2) engendre une famille de schémas itératifs qu'on appellera *famille de base*.

On a montré au chapitre V que pour l'étude de la convergence de la méthode itérative il faut rechercher le comportement dans H_D de la norme de l'erreur $z_k = y_k - u$ pour $k \to \infty$, où y_k est l'approximation itérative obtenue avec le schéma (2) et u la solution de l'équation (1). La méthode itérative converge dans H_D si la norme d'erreur z_k tend dans H_D vers zéro quand k tend vers l'infini.

Comme la vitesse de convergence dépend du choix des paramètres d'itération τ_k, ces derniers doivent être choisis de manière que la vitesse de convergence soit maximale.

2. Problème des erreurs. Etudions d'abord la convergence des schémas itératifs à deux couches (2). A cette fin on obtient l'équation à laquelle satisfait l'erreur z_k.

En posant $y_k = z_k + u$ pour $k = 0, 1, \ldots$ dans (2) et, compte tenu de l'équation (1), il vient

$$B \frac{z_{k+1} - z_k}{\tau_{k+1}} + Az_k = 0, \quad k = 0, 1, \ldots, \quad z_0 = y_0 - u,$$

autrement dit, l'erreur z_k satisfait à une équation homogène. En résolvant cette équation en z_{k+1}:

$$z_{k+1} = (E - \tau_{k+1} B^{-1} A)\, z_k$$

et en admettant que $z_k = D^{-1/2} x_k$, passons à l'équation pour l'erreur équivalente x_k, qui ne comprendra qu'un seul opérateur. L'équation pour x_k aura la forme

$$x_{k+1} = S_{k+1} x_k, \quad S_{k+1} = E - \tau_{k+1} C, \quad k = 0, 1, \ldots, \tag{3}$$

où $C = D^{1/2} B^{-1} A D^{-1/2}$. En vertu de la substitution effectuée, se vérifie l'égalité

$$\| x_k \| = \| D^{1/2} z_k \| = \| z_k \|_D,$$

aussi le problème de l'étude de la convergence de la méthode itérative (2) dans H_D se réduit-il à la recherche de la suite numérique $\| x_k \|$, $k = 1, 2, \ldots$, où x_k est défini dans (3).

Cherchons la solution de l'équation (3). De (3) il vient

$$x_k = T_{k,0} x_0, \quad T_{k,0} = \prod_{i=1}^{k} S_i = S_n S_{n-1} \ldots S_1.$$

De là s'ensuit l'estimation suivante pour la norme d'erreur z_k dans H_D:

$$\| z_k \|_D = \| x_k \| \leqslant \| T_{k,0} \| \| x_0 \| = \| T_{k,0} \| \| z_0 \|_D. \qquad (4)$$

L'opérateur $T_{k,0}$ est dit *opérateur résolvant* de la k-ième itération, tandis que S_k est l'*opérateur de passage* de la $(k-1)$-ième itération à la k-ième.

Il s'ensuit de l'estimation (4) que la méthode itérative (2) converge dans H_D si la norme de l'opérateur résolvant $T_{k,0}$ tend vers zéro quand k tend vers l'infini.

Ainsi le problème de l'étude de la convergence de la méthode itérative (2) dans H_D se réduit-il à la recherche du comportement de la norme de l'opérateur résolvant $T_{k,0}$ dans l'espace H en fonction du numéro d'itération k.

L'opérateur résolvant $T_{k,0}$ est défini par l'opérateur C et les paramètres d'itération $\tau_1, \tau_2, \ldots, \tau_k$.

En admettant l'opérateur C fixé, posons le problème du choix des paramètres $\{\tau_k\}$ de manière que la méthode itérative converge. Parmi les méthodes itératives convergentes la méthode *optimale* sera apparemment celle dont les paramètres $\{\tau_k\}$ garantissent l'acquisition de la précision voulue $\varepsilon > 0$ en un nombre minimal d'itérations. En vertu de l'estimation (4), on peut donner à cette exigence la forme équivalente suivante: construire pour un n donné le jeu de paramètres itératifs $\tau_1, \tau_2, \ldots, \tau_n$ pour lequel la norme de l'opérateur $T_{n,0}$ soit minimale.

3. Cas d'opérateur autoadjoint. Posons maintenant de façon très stricte le problème du meilleur choix des paramètres d'itération pour le schéma à deux couches (2). Ce problème présentera une solution, si des hypothèses bien déterminées seront faites relativement aux opérateurs A, B et D. Formulons ces hypothèses.

1) Posons que les opérateurs A, B et D sont tels que l'opérateur $DB^{-1}A$ est autoadjoint dans H. Si cette hypothèse est vérifiée, on dira qu'on est dans le cas d'opérateurs autoadjoints.

2) Soient γ_1 et γ_2 les constantes de l'équivalence énergétique des opérateurs D et $DB^{-1}A$, c'est-à-dire les constantes des inégalités

$$\gamma_1 D \leqslant DB^{-1}A \leqslant \gamma_2 D, \quad \gamma_1 > 0, \quad DB^{-1}A = (DB^{-1}A)^*. \qquad (5)$$

La seconde hypothèse détermine le type de l'information à priori sur les opérateurs du schéma itératif; cette information est utilisée pour l'établissement des formules pour les paramètres d'itération dans le cas d'opérateurs autoadjoints. L'exemple le plus simple, où l'hypothèse de l'opérateur $DB^{-1}A$ autoadjoint est satisfaite, est le suivant: $A = A^*$, $D = B = E$, c'est-à-dire on est en présence d'un schéma explicite dans l'espace initial H pour l'équation (1) à opérateur A autoadjoint. Dans ce cas l'information à priori se réduit à

la fixation des bornes de l'opérateur A. Des exemples plus compliqués du choix de l'opérateur D seront étudiés plus loin.

Supposons donc que les conditions (5) sont remplies. De (5) il s'ensuit que l'opérateur $C = D^{-1/2}(DB^{-1}A)D^{-1/2}$ est autoadjoint dans H, quant à γ_1 et γ_2, ce sont ses bornes, autrement dit

$$\gamma_1 E \leqslant C \leqslant \gamma_2 E, \quad \gamma_1 > 0, \quad C = C^* = D^{-1/2}(DB^{-1}A)D^{-1/2}. \quad (6)$$

En effet, en posant dans les inégalités

$$\gamma_1 (Dx, x) \leqslant (DB^{-1}Ax, x) \leqslant \gamma_2 (Dx, x)$$

$x = D^{-1/2}y$, on obtient les inégalités (6). Par conséquent, les hypothèses formulées plus haut sur les opérateurs A, B et D sont équivalentes aux conditions (6).

Formulons maintenant le problème du choix optimal des paramètres d'itération pour le schéma (2). De la définition de l'opérateur résolvant $T_{k,0}$ et des conditions (6) il s'ensuit que l'opérateur $T_{k,0} = T_{k,0}(C)$ est autoadjoint dans H et la norme du polynôme opératoriel $T_{n,0}(C)$ s'estime de la façon suivante:

$$\| T_{n,0} \| \leqslant \max_{\gamma_1 \leqslant t \leqslant \gamma_2} \left| \prod_{k=1}^{n} (1 - \tau_k t) \right|.$$

A partir de l'estimation (4) on tire que dans le cas d'opérateurs autoadjoints les paramètres d'itération $\tau_1, \tau_2, \ldots, \tau_n$ doivent être choisis de façon que le maximum du module du polynôme $P_n(t) = \prod_{k=1}^{n} (1 - \tau_k t)$, construit en fonction de ces paramètres, soit minimal sur le tronçon $[\gamma_1, \gamma_2]$, c'est-à-dire qu'il faut trouver les paramètres en partant des conditions

$$\min_{\{\tau_k\}} \max_{\gamma_1 \leqslant t \leqslant \gamma_2} \left| \prod_{k=1}^{n} (1 - \tau_k t) \right| = \max_{\gamma_1 \leqslant t \leqslant \gamma_2} | P_n(t) |.$$

Alors pour l'erreur de la méthode (2) se vérifiera l'estimation $\| z_n \|_D \leqslant q_n \| z_0 \|_D$, où

$$q_n = \max_{\gamma_1 \leqslant t \leqslant \gamma_2} | P_n(t) |.$$

Le problème formulé plus haut est le problème classique du minimax. On donnera au § 2 la solution de ce problème et on fournira le jeu des paramètres d'itération $\tau_1, \tau_2, \ldots, \tau_n$. La méthode itérative avec un tel jeu de paramètres est appelée *méthode de Tchébychev*. Dans la littérature spécialisée cette méthode est également dénommée *méthode de Richardson*.

§ 2. Méthode de Tchébychev à deux couches

1. Construction d'un jeu de paramètres d'itération. Au § 1 on a montré que la construction d'un jeu optimal de paramètres d'itération $\tau_1, \tau_2, \ldots, \tau_n$ se réduit à la recherche du polynôme $P_n(t)$ de

l'aspect $P_n(t) = \prod_{k=1}^{n}(1 - \tau_k t)$ dont le maximum du module sur le segment $[\gamma_1, \gamma_2]$ est minimal.

Résolvons ce problème. Vu que la forme du polynôme est déterminée par la condition de normalisation $P_n(0) = 1$, le problème posé se formule de la façon suivante: parmi tous les polynômes de degré n prenant au point $t = 0$ la valeur 1 trouver le polynôme s'écartant le moins de zéro sur le segment $[\gamma_1, \gamma_2]$ ne comprenant pas le point 0.

La solution de ce problème a été obtenue par le mathématicien russe V. A. Markov en 1892 et est donnée dans l'annexe. Le polynôme cherché $P_n(t)$ a la forme

$$P_n(t) \equiv q_n T_n\left(\frac{1-\tau_0 t}{\rho_0}\right), \quad q_n = \frac{1}{T_n\left(\frac{1}{\rho_0}\right)}, \tag{1}$$

où $T_n(x)$ est le polynôme de Tchébychev de première espèce de degré n,

$$T_n(x) = \begin{cases} \cos(n \arccos x), & |x| \leqslant 1, \\ \operatorname{ch}(n \operatorname{Arch} x), & |x| \geqslant 1, \end{cases}$$

$$q_n = \frac{2\rho_1^n}{1+\rho_1^{2n}}, \quad \tau_0 = \frac{2}{\gamma_1+\gamma_2}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}. \tag{2}$$

En outre, $\max_{\gamma_1 \leqslant t \leqslant \gamma_2} |P_n(t)| = q_n$. De là s'ensuit l'estimation pour la norme d'erreur z_n dans H_D:

$$\| z_n \|_D \leqslant q_n \| z_0 \|_D, \tag{3}$$

où q_n est défini dans (2).

Cherchons les formules des paramètres d'itération. Vu que les polynômes des premier et second membres de (1) prennent la même valeur égale à 1 pour $t = 0$, l'identité dans (1) ne se réalisera que dans le cas où les ensembles des racines des polynômes $P_n(t)$ et $T_n\left(\frac{1-\tau_0 t}{\rho_0}\right)$ coïncideront. Le polynôme $P_n(t)$ possède les racines $1/\tau_k$, $k = 1, 2, \ldots, n$, tandis que le polynôme $T_n(x)$ a des racines égales à $-\cos\left(\frac{2i-1}{2n}\pi\right)$, $i = 1, 2, \ldots, n$. Si l'on désigne par $\mathfrak{M}_n$ l'ensemble des racines du polynôme de Tchébychev $T_n(x)$:

$$\mathfrak{M}_n = \left\{-\cos\frac{2i-1}{2n}\pi, \quad i = 1, 2, \ldots, n\right\}, \tag{4}$$

on obtiendra la formule suivante pour les paramètres d'itération:

$$\tau_k = \tau_0/(1 + \rho_0\mu_k), \quad \mu_k \in \mathfrak{M}_n, \quad k = 1, 2, \ldots, n. \tag{5}$$

$\mu_k \in \mathfrak{M}_n$ signifie qu'en guise de μ_k on doit choisir successivement tous les éléments de l'ensemble $\mathfrak{M}_n$.

A partir de la formule ainsi obtenue pour les paramètres τ_k on déduit que pour le calcul des paramètres d'itération il faut fixer le nombre d'itérations n. Aussi passons à l'appréciation du nombre d'itérations. Habituellement, pour condition d'achèvement du processus d'itérations on choisit l'inégalité

$$\| z_n \|_D \leqslant \varepsilon \| z_0 \|_D$$

en désignant pour *nombre d'itérations* le plus petit nombre n pour lequel l'inégalité est satisfaite.

Il s'ensuit de (3) que pour la méthode étudiée le nombre d'itérations se déduit de l'inégalité $q_n \leqslant \varepsilon$. En recourant à (2), résolvons cette inégalité. Il vient

$$n \geqslant n_0(\varepsilon), \quad n_0(\varepsilon) = \ln\left(\frac{1}{\varepsilon} + \sqrt{\frac{1}{\varepsilon^2} - 1}\right) \Big/ \ln \frac{1}{\rho_1}.$$

On utilise généralement une formule plus simple pour $n_0(\varepsilon)$

$$n \geqslant n_0(\varepsilon), \quad n_0(\varepsilon) = \ln \frac{2}{\varepsilon} \Big/ \ln \frac{1}{\rho_1}. \qquad (6)$$

Après avoir trouvé le nombre d'itérations exigé n, on peut construire, suivant la formule (5), le jeu des paramètres d'itération.

Bref, pour un schéma implicite à deux couches

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + A y_k = f, \quad k = 0, 1, \ldots, y_0 \in H, \qquad (7)$$

on a démontré le

Théorème 1. *Soient remplies les conditions*

$$\gamma_1 D \leqslant DB^{-1}A \leqslant \gamma_2 D, \ \gamma_1 > 0, \ DB^{-1}A = (DB^{-1}A)^*, \ D = D^* > 0. \qquad (8)$$

Alors le processus d'itération de Tchébychev (7), (4), (5), (2) *converge dans* H_D, *et pour l'erreur* z_n *on a l'estimation* (3). *Pour le nombre d'itérations l'estimation* (6) *est vraie.*

Les estimations obtenues permettent de conclure que dans le cas d'opérateurs autoadjoints la vitesse de convergence de la méthode de Tchébychev est fonction du rapport $\xi = \gamma_1/\gamma_2$, cette vitesse étant d'autant plus élevée que ξ est plus grand.

2. Impossibilité d'améliorer l'estimation à priori. Montrons maintenant que dans la classe des approximations initiales quelconques y_0 l'estimation de l'erreur de la méthode de Tchébychev, donnée dans le théorème 1, ne peut être améliorée dans le cas d'un espace H de dimension finie. Il suffit d'indiquer l'approximation initiale y_0 pour laquelle avec une norme d'erreur équivalente x_k on a l'égalité $\| x_n \| = q_n \| x_0 \|$. On cherchera l'erreur initiale x_0 vérifiant cette

égalité, quant à l'approximation initiale y_0, en vertu du rapport entre les erreurs z_k et x_k, $z_k = D^{-1/2}x_k$, elle sera déterminée suivant la formule $y_0 = u + D^{-1/2}x_0$.

Cherchons l'inconnue x_0. Soit H un espace de dimension finie ($H = H_N$). Vu que l'opérateur C est autoadjoint dans H, il existe un système complet de fonctions propres $v_1, v_2, \ldots, v_N$ de l'opérateur C. Désignons par λ_k la valeur propre de l'opérateur C associée à la fonction propre v_k. Posons que les valeurs propres sont ordonnées $\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_N$. En qualité de bornes de l'opérateur C on peut alors prendre $\gamma_1 = \lambda_1$ et $\gamma_2 = \lambda_N$.

En guise d'erreur initiale x_0 choisissons la fonction propre v_1. A partir de l'équation pour l'erreur x_k:

$$x_{k+1} = (E - \tau_{k+1}C)\, x_k, \quad k = 0, 1, \ldots, \quad x_0 = v_1$$

et l'égalité $Cv_k = \lambda_k v_k$, on obtient successivement

$$x_1 = (E - \tau_1 C)\, x_0 = (1 - \tau_1\gamma_1)\, v_1 = (1 - \tau_1\gamma_1)\, x_0,$$

$$x_2 = (E - \tau_2 C)\, x_1 = (1 - \tau_1\gamma_1)\,(E - \tau_2 C)\, x_0 = (1 - \tau_1\gamma_1)\,(1 - \tau_2\gamma_1)\, x_0,$$

. . . .

$$x_n = \prod_{k=1}^{n} (1 - \tau_k\gamma_1)\, x_0 = P_n(\gamma_1)(x_0).$$

En portant dans (1) $t = \gamma_1$ et compte tenu de l'égalité $1 - \tau_0\gamma_1 = \rho_0$, calculons $P_n(\gamma_1) = q_n T_n(1) = q_n$ et, par suite,

$$x_n = q_n x_0, \quad \| x_n \| = q_n \| x_0 \|,$$

ce qu'il fallait démontrer.

Bref, on a montré que l'estimation à priori obtenue dans le théorème 1 ne peut être améliorée dans la classe des approximations initiales arbitraires.

3. Exemples de choix de l'opérateur D. Donnons quelques exemples de choix de l'opérateur D. Rappelons que la méthode de Tchébychev est considérée dans l'hypothèse de l'opérateur $DB^{-1}A$ autoadjoint. On indiquera plus loin les conditions imposées aux opérateurs A et B pour que cette hypothèse soit vérifiée une fois l'opérateur D choisi. Pour chaque choix concret de l'opérateur D on indiquera les inégalités fixant l'information à priori sur les opérateurs du schéma itératif. Cette information est utilisée pour la construction du jeu de paramètres d'itération dans la méthode de Tchébychev.

V o y o n s l e p r e m i e r e x e m p l e. Soient A et B les opérateurs autoadjoints et définis positifs dans H. En guise d'opérateur D on peut alors choisir l'un des opérateurs suivants: A ou B. Si, de plus, l'opérateur B est borné dans H, on peut prendre $D = AB^{-1}A$. Dans ce cas l'information à priori se réduit à la fixation des constantes de l'équivalence énergétique des opérateurs A et B:

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \quad \gamma_1 > 0, \quad B > 0. \tag{9}$$

En effet, il faut montrer que les conditions suivantes sont remplies: l'opérateur D choisi est autoadjoint et défini positif dans H, l'opérateur $DB^{-1}A$ est autoadjoint dans H, les inégalités (8) et (9) étant équivalentes.

Le fait que les opérateurs D et $DB^{-1}A$ sont autoadjoints dans tous les cas étudiés est la conséquence de ce que les opérateurs A et B sont aussi autoadjoints. Au cas où $D = A$ ou $D = B$ la définissabilité positive de D découle de celle de A et B. Montrons maintenant que l'opérateur $D = AB^{-1}A$ est également défini positif dans H.

De fait, supposons que sont remplies les conditions formulées plus haut relatives aux opérateurs A et B: $A = A^* \geqslant \alpha E$, $B = B^* \geqslant \beta E$, $\|Bx\| \leqslant M \|x\|$, $\alpha, \beta > 0$, $M < \infty$. A partir de ces conditions et des lemmes 6 et 8 du § 1, ch. V, on tire que $B^{-1} \geqslant \frac{1}{M} E$ et $(Ax, Ax) \geqslant \alpha (Ax, x) \geqslant \alpha^2 (x, x)$. De là on obtient pour l'énergie de l'opérateur D l'estimation par le bas

$$(Dx, x) = (AB^{-1}Ax, x) = (B^{-1}Ax, Ax) \geqslant$$

$$\geqslant \frac{1}{M} (Ax, Ax) \geqslant \frac{\alpha^2}{M} (x, x), \quad \text{c'est-à-dire } D \geqslant \frac{\alpha^2}{M} E.$$

Par conséquent, la définissabilité positive de l'opérateur $D = AB^{-1}A$ est démontrée.

Montrons maintenant que les inégalités (8) et (9) sont équivalentes dans l'exemple examiné. En effet, supposons vérifiées les inégalités (9):

$$\gamma_1 (Bx, x) \leqslant (Ax, x) \leqslant \gamma_2 (Bx, x), \quad \gamma_1 > 0. \tag{10}$$

Si $D = B$, $DB^{-1}A = A$, les inégalités (10) et (8) coïncident de même. Soit maintenant $D = AB^{-1}A$. Dans ce cas $DB^{-1}A = AB^{-1} \times AB^{-1}A$ et, posant dans (10) $x = B^{-1}Ay$, il vient

$$\gamma_1 (AB^{-1}Ay, y) \leqslant (AB^{-1}Ay, B^{-1}Ay) \leqslant \gamma_2 (AB^{-1}Ay, y)$$

ou

$$\gamma_1 (Dy, y) \leqslant (DB^{-1}Ay, y) \leqslant \gamma_2 (Dy, y),$$

c'est-à-dire que l'on obtient les inégalités (8). Le passage inverse de (8) à (10) est évident.

Soit $D = A$, alors $DB^{-1}A = AB^{-1}A$. Il s'ensuit du lemme 9 du § 1, ch. V, que pour les opérateurs A et B autoadjoints et définis positifs les inégalités (10) et les inégalités

$$\gamma_1 (A^{-1}x, x) \leqslant (B^{-1}x, x) \leqslant \gamma_2 (A^{-1}x, x), \quad \gamma_1 > 0$$

sont équivalentes. En posant ici $x = Ay$, on aboutit à l'inégalité (8). Le passage inverse est évident.

Cette inégalité permet de démontrer aussitôt que D est défini positif:

$$(Dx, x) \geqslant \alpha\gamma_1 (x, x).$$

En effet, $(Dx, x) = (B^{-1}Ax, Ax) \geqslant \gamma_1 (A^{-1}Ax, Ax) = \gamma_1 (Ax, x) \geqslant \geqslant \gamma_1 \alpha (x, x)$.

Second exemple. Posons que les opérateurs A et B sont autoadjoints et définis positifs dans H ainsi que permutables: $A = A^* > 0$, $B = B^* > 0$, $AB = BA$. Si en guise d'opérateur D on choisit l'opérateur A^2, l'information à priori sera alors fixée sous forme des inégalités (9).

En effet, le fait que l'opérateur D est autoadjoint et défini positif est la conséquence de ce que l'opérateur A est autoadjoint non dégénéré. Ensuite, $DB^{-1}A = A (AB^{-1}) A$ et puisque les opérateurs A et B sont permutables, les opérateurs A et B^{-1} sont donc aussi permutables. De là, une fois les opérateurs A et B autoadjoints, il s'ensuit que l'opérateur $DB^{-1}A$ est aussi autoadjoint.

Les inégalités (8) prennent dans ce cas la forme

$$\gamma_1 (Ax, Ax) \leqslant (AB^{-1}Ax, Ax) \leqslant \gamma_2 (Ax, Ax), \quad \gamma_1 > 0.$$

En posant ici $x = A^{-1}B^{1/2}y$ et utilisant la permutabilité de la racine de l'opérateur B avec l'opérateur A, il vient

$$\gamma_1 (By, y) \leqslant (Ay, y) \leqslant \gamma_2 (By, y),$$

autrement dit, on obtient l'inégalité (9). Le passage inverse de (9) à (8) est évident.

Voyons encore un exemple. Soient A et B des opérateurs quelconques non dégénérés satisfaisant à la condition

$$B^*A = A^*B. \tag{11}$$

Si en guise de D on choisit l'opérateur A^*A, l'information à priori peut être donnée sous forme des inégalités

$$\gamma_1 (Bx, Bx) \leqslant (Ax, Bx) \leqslant \gamma_2 (Bx, Bx), \quad \gamma_1 > 0. \tag{12}$$

L'opérateur D est évidemment autoadjoint et il est défini positif vu la non-dégénérescence de l'opérateur A. L'opérateur B étant non dégénéré, les conditions (11) peuvent se transcrire sous forme des conditions $AB^{-1} = (B^*)^{-1} A^*$ qui expriment que l'opérateur AB^{-1} est autoadjoint. De là on obtient que l'opérateur $DB^{-1}A = A^*AB^{-1}A$ est autoadjoint dans H. Ensuite, en posant dans (12) $x = B^{-1}Ay$, il vient

$$\gamma_1 (Ay, Ay) \leqslant (AB^{-1}Ay, Ay) \leqslant \gamma_2 (Ay, Ay)$$

ou

$$\gamma_1 (Dy, y) \leqslant (DB^{-1}Ay, y) \leqslant \gamma_2 (Dy, y).$$

C'est ainsi que des inégalités (12) s'ensuivent les inégalités (8). Le passage inverse de (8) à (12) est évident.

Notons en conclusion qu'en cas d'opérateurs A et B autoadjoints définis positifs et bornés dans H la méthode itérative de Tchébychev converge dans H_D, où $D = A$, B ou $AB^{-1}A$ (et, si, en outre, A et B

sont permutables, également pour $D = A^2$), à la même vitesse définie par le rapport des constantes γ_1 et γ_2 des inégalités (9).

Les cas de $D = AB^{-1}A$ et de $D = A^*A$ méritent une attention particulière. Avec ce choix de l'opérateur D la norme de l'erreur dans H_D peut être calculée au cours des itérations. En effet, avec $D = AB^{-1}A$, il vient

$$\| z_n \|_D^2 = (Dz_n, z_n) = (B^{-1}Az_n, Az_n) = (B^{-1}r_n, r_n) = (w_n, r_n),$$

et avec $D = A^*A$:

$$\| z_n \|_D^2 = (Az_n, Az_n) = (r_n, r_n),$$

où $r_n = Az_n = Ay_n - Au = Ay_n - f$ est le résidu de la n-ième itération et $w_n = B^{-1}r_n$ la correction. Ces grandeurs peuvent être obtenues au cours des itérations.

4. Stabilité de la méthode sous le rapport des calculs. En étudiant la convergence de la méthode de Tchébychev on a admis que le processus de calcul était parfait, c'est-à-dire que les calculs s'effectuaient avec un nombre infini de chiffres. Dans un calcul réel toutes les opérations de calcul se réalisent avec un nombre fini de chiffres et à chaque étape du calcul apparaissent des erreurs d'arrondi. Les erreurs d'arrondi associées aux opérations arithmétiques engendrent l'erreur de la méthode.

Dans les méthodes itératives, l'erreur de calcul de la méthode est constituée des erreurs impliquées par chaque itération. Si le nombre d'itérations est suffisamment grand et la méthode itérative est susceptible d'accumuler les erreurs d'arrondi de chaque itération, l'erreur de calcul d'une telle méthode peut s'avérer si grande qu'on aboutit à une perte totale de précision, et l'approximation itérative y_n différera fortement de la solution cherchée. Aussi pour les méthodes itératives est-il important d'étudier le mécanisme de formation des erreurs de calcul et de déceler les stades de l'algorithme où se produit l'accroissement de l'erreur de calcul de la méthode. Dans nombre de cas certaines modifications du processus de calcul permettent d'atténuer sensiblement l'accroissement de l'erreur de calcul et de rendre la méthode applicable à des utilisations pratiques.

Une autre particularité du processus réel de calcul est en relation avec l'existence pour l'ordinateur d'un « zéro » et d'un « infini machine ». Ces notions caractérisent l'ordre admissible de nombres pouvant être introduits dans l'ordinateur. Par exemple, dans l'ordinateur BESM-6, en régime de précision unique, il peut être introduit des nombres réels dont la valeur absolue appartient à la gamme allant de 10^{-19} à 10^{19}. Ce sont justement les limites du « zéro » et de l'« infini machine ». Si les calculs sur ordinateur aboutissent à une valeur sortant de cet intervalle, les calculs s'arrêtent et il y a lieu à « arrêt d'urgence ». Aussi l'exigence envers la « continuité de service » du processus itératif est-elle toute naturelle.

Bref, les méthodes itératives doivent assurer la « continuité de service » (de l'ordinateur) et présenter une stabilité par rapport aux erreurs d'arrondi.

On a construit au point 1, § 2, la méthode de Tchébychev à deux couches. Le théorème 1 montre qu'après n itérations aux paramètres $\tau_k = \tau_0/(1 + \rho_0\mu_k)$, $\mu_k \in \mathfrak{M}_n$, $k = 1, 2, \ldots, n$, l'estimation $\| z_n \|_D \leqslant q_n \| z_0 \|_D$ sera vérifiée pour l'erreur z_n. En guise de μ_k on choisit successivement tous les éléments de l'ensemble $\mathfrak{M}_n$, l'ordre suivi étant quelconque.

Etudions la stabilité des calculs de la méthode de Tchébychev. Pour être plus concret, admettons que μ_k est le k-ième élément de l'ensemble $\mathfrak{M}_n$. Dans ce cas les différents ensembles $\mathfrak{M}_n$ ordonnés engendreront des différentes suites $\{\mu_k\}$ et, partant, des différentes suites de paramètres d'itération $\{\tau_k\}$.

Du point de vue d'un processus de calcul idéal, toutes les suites de paramètres d'itération de Tchébychev sont équivalentes, c'est-à-dire que chaque suite doit aboutir à une même approximation y_n et, partant, à une même précision après exécution de n itérations. L'apparition d'erreurs d'arrondi dans les calculs réels implique la non-équivalence des suites de paramètres d'itération.

Illustrons cette assertion par un exemple. Supposons que sur le maillage $\bar\omega = \{x_i = ih,\ 0 \leqslant i \leqslant N,\ h = 1/N\}$ il s'agit de trouver la solution du problème de différences suivant:

$$\Lambda y = y_{\bar{x}x} - dy = -\varphi(x), \quad x \in \omega,$$

$$y(0) = 0, \quad y(1) = 1, \quad d = \text{const} > 0.$$

Au § 2, ch. V, il a été montré que le problème de différences peut être réduit à l'équation opératorielle

$$Ay = f, \tag{13}$$

dont l'opérateur A se détermine de la façon suivante: $Ay = -\Lambda\mathring{y}$, où $y \in H$, $\mathring{y} \in \mathring{H}$, $\mathring{y}(x) = y(x)$ pour $x \in \omega$. $\mathring{H}$ est ici un ensemble des fonctions de mailles associées à $\bar\omega$ et s'annulant pour $x = 0$ et $x = 1$, quant à H, c'est l'espace de fonctions de mailles données sur ω avec produit scalaire $(u, v) = \sum_{x\in\omega} u(x)\, v(x)\, h$. Le second membre f de l'équation (13) ne diffère du second membre φ du schéma aux différences qu'aux nœuds frontières du maillage: $f(x) = \varphi(x)$, $h \leqslant x \leqslant 1 - 2h$, $f(1 - h) = \varphi(1 - h) + 1/h^2$.

Pour la résolution approchée de l'équation (13), recourrons à la méthode explicite de Tchébychev

$$\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \ldots, y_0 \in H. \tag{14}$$

Comme les opérateurs A et $B = E$ sont autoadjoints dans H, il s'ensuit des exemples examinés au point 3, § 2, qu'en guise d'infor-

mation à priori pour la méthode de Tchébychev (14) il suffit de fixer les bornes de l'opérateur A : $\gamma_1 E \leqslant A \leqslant \gamma_2 E$, $\gamma_1 > 0$, si en qualité de l'opérateur D est choisi l'opérateur $B = E$. γ_1 et γ_2 coïncident, apparemment, avec les valeurs propres minimale et maximale de l'opérateur de différence Λ, c'est-à-dire

$$\gamma_1 = \frac{4}{h^2} \sin^2 \frac{\pi h}{2} + d, \quad \gamma_2 = \frac{4}{h^2} \cos^2 \frac{\pi h}{2} + d.$$

Les paramètres d'itération τ_k se calculent suivant les formules

$$\tau_k = \tau_0/(1 + \rho_0 \mu_k), \quad \mu_k \in \mathfrak{M}_n, \quad k = 1, 2, \ldots, n,$$
$$\tau_0 = 2/(\gamma_1 + \gamma_2), \quad \rho_0 = (\gamma_2 - \gamma_1)/(\gamma_2 + \gamma_1). \qquad (15)$$

On a examiné trois suites de paramètres d'itération définies par les $\mathfrak{M}_n$ ordonnés suivants:

1) suite « directe »

$$\mathfrak{M}_n = \mathfrak{M}_n^{(1)} = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}, \text{ c'est-à-dire } \mu_k = \sigma_k, \quad k = 1, 2, \ldots, n;$$

2) suite « inverse »

$$\mathfrak{M}_n = \mathfrak{M}_n^{(2)} = \{\sigma_n, \sigma_{n-1}, \ldots, \sigma_1\}, \text{ c'est-à-dire } \mu_k = \sigma_{n-k+1}, \quad k = 1, 2, \ldots, n;$$

3) suite « alternée »

$$\mathfrak{M}_n = \mathfrak{M}_n^{(3)} = \{\sigma_1, \sigma_n, \sigma_2, \sigma_{n-1}, \ldots\}, \text{ c'est-à-dire } \mu_{2k-1} = \sigma_k, \quad \mu_{2k} = \sigma_{n-k+1}, \quad k = 1, 2, \ldots, n/2.$$

On a posé ici $\sigma_k = -\cos \frac{2k-1}{2n} \pi$.

Les calculs s'effectuaient de la façon suivante: on fixait le nombre d'itérations n et, suivant le schéma (14), (15), pour chaque suite de paramètres d'itération on effectuait n itérations. La précision réelle à laquelle on aboutissait après n itérations s'évaluait par la formule

$$\varepsilon_{\text{réel}} = \frac{\| y_n - u \|}{\| y_0 - u \|}.$$

A titre de comparaison on calculait la quantité q_n, où

$$q_n = \frac{2\rho_1^n}{1 + \rho_1^{2n}}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2},$$

définissant la précision théorique de la méthode lorsque le nombre d'itérations est n. Dans tous les calculs l'approximation initiale y_0 était prise égale à zéro sur ω. La solution précise du problème de différences $y(x) = x$ correspond au second membre $\varphi(x) = dx$. Le coefficient d était choisi de la sorte que γ_1 fût égal à 0,1 :

$$\gamma_1 = 0{,}1, \quad \gamma_2 = 0{,}1 + \frac{4}{h^2} \cos \pi h, \quad \frac{1}{\xi} = \frac{40}{h^2} \cos \pi h + 1.$$

Tableau 5

n	q_n	$e_{\text{réel}}$			
		$\mathfrak{M}_n^{(1)}$	$\mathfrak{M}_n^{(2)}$	$\mathfrak{M}_n^{(3)}$	$\mathfrak{M}_n^{*}$
16	$8{,}79\cdot10^{-1}$	$8{,}14\cdot10^{-1}$	$8{,}14\cdot10^{-1}$	$8{,}14\cdot10^{-1}$	$8{,}14\cdot10^{-1}$
24	$7{,}58\cdot10^{-1}$	$9{,}62\cdot10^{-1}$	$7{,}11\cdot10^{-1}$	$7{,}11\cdot10^{-1}$	$7{,}11\cdot10^{-1}$
32	$6{,}30\cdot10^{-1}$	$3{,}38\cdot10^{3}$	$3{,}55\cdot10^{2}$	$5{,}63\cdot10^{-1}$	$5{,}63\cdot10^{-1}$
40	$5{,}09\cdot10^{-1}$	$3{,}07\cdot10^{7}$	$2{,}44\cdot10^{6}$	$5{,}03\cdot10^{-1}$	$4{,}85\cdot10^{-1}$
48	$4{,}04\cdot10^{-1}$	arrêt	$3{,}46\cdot10^{10}$	$2{,}47\cdot10^{0}$	$3{,}64\cdot10^{-1}$
56	$3{,}17\cdot10^{-1}$	—	$1{,}02\cdot10^{15}$	$2{,}29\cdot10^{2}$	$3{,}10\cdot10^{-1}$
64	$2{,}47\cdot10^{-1}$	—	arrêt	$1{,}87\cdot10^{4}$	$2{,}23\cdot10^{-1}$
72	$1{,}92\cdot10^{-1}$	—	—	$1{,}73\cdot10^{6}$	$1{,}72\cdot10^{-1}$
80	$1{,}49\cdot10^{-1}$	—	—	arrêt	$1{,}44\cdot10^{-1}$
...	...	...	...	...	...
256	$4{,}97\cdot10^{-4}$	—	—	—	$4{,}80\cdot10^{-4}$
...	...	...	...	...	...
512	$1{,}23\cdot10^{-7}$	—	—	—	$1{,}15\cdot10^{-7}$

Les résultats des calculs pour $N = 10$ sont donnés au tableau 5. Dans ce tableau, outre les suites mentionnées de paramètres d'itération, sont indiqués les résultats pour l'ensemble $\mathfrak{M}_n^*$ maximalement ordonné qui sera décrit plus loin.

Les calculs effectués montrent que dans le processus réel de calculs les suites étudiées de paramètres d'itération ne sont pas équivalentes. Les calculs ont fait ressortir deux particularités caractéristiques de ce processus réel: possibilité d'« arrêt de la machine » dû à l'accroissement de solutions itératives intermédiaires et possibilité de perte de précision finale en cas de continuité de service à cause de l'accumulation d'erreurs d'arrondi.

La raison de cette instabilité de calcul de la méthode pour certaines suites de paramètres d'itération réside dans le fait que la norme de l'opérateur de passage d'une itération à l'autre $S_k = E - \tau_k C$ est plus grande que l'unité pour certaines valeurs de k.

En effet, puisque S est un opérateur autoadjoint dans H, $\| S_k \| = \sup\limits_{\|x\|=1} |(S_k x, x)|$. En utilisant les bornes γ_1 et γ_2 de l'opérateur C

$$\gamma_1 E \leqslant C \leqslant \gamma_2 E, \quad \gamma_1 > 0,$$

on obtient

$$(1 - \tau_k\gamma_2)\, E \leqslant E - \tau_k C \leqslant (1 - \tau_k\gamma_1)\, E.$$

Portons-y τ_k tiré de (15) et tenons compte de l'égalité $1 - \rho_0 = \tau_0\gamma_1$, $1 + \rho_0 = \tau_0\gamma_2$. Il vient

$$-\frac{\rho_0(1-\mu_k)}{1+\rho_0\mu_k}\, E \leqslant S_k \leqslant \frac{\rho_0(1+\mu_k)}{1+\rho_0\mu_k}\, E$$

et, par conséquent,

$$\|S_k\| = \begin{cases} \dfrac{\rho_0(1+\mu_k)}{1+\rho_0\mu_k} < 1, & \mu_k \geqslant 0, \\ \dfrac{\rho_0(1-\mu_k)}{1+\rho_0\mu_k}, & \mu_k < 0. \end{cases}$$

Il s'ensuit de là que $\|S_k\| > 1$ pour $\mu_k < -(1-\rho_0)/(2\rho_0)$. Comme $\mu_k \in \mathfrak{M}_n$, on a

$$-\cos\frac{\pi}{2n} \leqslant \mu_k \leqslant -\cos\frac{2n-1}{2n}\pi = \cos\frac{\pi}{2n}, \quad k = 1, 2, \ldots, n,$$

et, par suite, pour un grand nombre de numéros k la norme $\|S_k\| > 1$ (le nombre de ces numéros de k est environ égal à $n/2$). Aussi si l'on utilise successivement un trop grand nombre de paramètres τ_k pour lesquels la norme de l'opérateur S_k est supérieure à l'unité, il peut se produire une accumulation d'erreurs de calcul et, partant, un accroissement d'approximations itératives impliquant l'instabilité de calcul de la méthode.

Le théorème 1 traduit en fait l'instabilité du schéma itératif sur la base des données initiales. Dans le cas de calculs réels il est nécessaire d'étudier la stabilité du schéma itératif également par rapport au second membre, car les erreurs d'arrondi peuvent être traitées comme des perturbations du second membre du schéma itératif à chaque itération.

Si l'on tient compte de l'erreur d'arrondi, au lieu d'une équation homogène pour l'erreur équivalente x_k on obtient une équation inhomogène

$$x_{k+1} = S_{k+1}x_k + \tau_{k+1}\varphi_{k+1}, \quad k = 0, 1, \ldots \tag{16}$$

$x_k = D^{1/2}(\bar{y}_k - u)$, où $\bar{y}_k$ est l'approximation itérative réelle.

En résolvant l'équation (16), on trouve $x_n = T_{n,0}x_0 + \sum_{j=1}^{n} \tau_j T_{n,j}\varphi_j$, où $T_{n,j} = \prod_{i=j+1}^{n} S_i$, $T_{n,n} = E$. On en tire l'estimation suivante:

$$\|x_n\| \leqslant \|T_{n,0}\|\,\|x_0\| + \sum_{j=1}^{n} \tau_j \|T_{n,j}\| \max_{1\leqslant j\leqslant n} \|\varphi_j\|. \tag{17}$$

L'estimation de la norme de l'opérateur $T_{n,0}$ est indépendante de la mise en ordre de l'ensemble $\mathfrak{M}_n$ et pour toute suite de paramètres de Tchébychev τ_k on a $\|T_{n,0}\| \leqslant q_n$. L'estimation pour $\sum_{j=1}^{n} \tau_j \|T_{n,j}\|$ est fonction de la mise en ordre de l'ensemble $\mathfrak{M}_n$. Il découle de (17) que l'ensemble $\mathfrak{M}_n$ doit être ordonné de manière que la somme mentionnée prenne une valeur minimale.

Le lemme suivant indique la valeur minimale possible que peut prendre cette somme.

L e m m e 1. *Si γ_1 et γ_2 sont des bornes précises de l'opérateur C pour toute mise en ordre de l'ensemble $\mathfrak{M}_n$, on a l'estimation*

$$\sum_{j=1}^{n} \tau_j \| T_{n,j} \| \geqslant \frac{1-q_n}{\gamma_1}.$$

En effet, de la définition de l'opérateur $T_{n,j}$ il ressort que

$$\tau_j T_{n,j} = (T_{n,j} - T_{n,j-1}) C^{-1}, \quad \sum_{j=1}^{n} \tau_j T_{n,j} = (E - T_{n,0}) C^{-1}.$$

Comme

$$\| (E - T_{n,0}) C^{-1} \| = \Big\| \sum_{j=1}^{n} \tau_j T_{n,j} \Big\| \leqslant \sum_{j=1}^{n} \tau_j \| T_{n,j} \|,$$

il suffit d'apprécier la norme de l'opérateur $(E - T_{n,0}) C^{-1}$. Cet opérateur est autoadjoint dans H et si γ_1 et γ_2 sont les bornes de l'opérateur C, on a

$$\| (E - T_{n,0}) C^{-1} \| \leqslant \max_{\gamma_1 \leqslant t \leqslant \gamma_2} \left| \frac{1 - q_n T_n \left(\frac{1 - \tau_0 t}{\rho_0} \right)}{t} \right| =$$

$$= \frac{1 - q_n T_n \left(\frac{1 - \tau_0 \gamma_1}{\rho_0} \right)}{\gamma_1} = \frac{1 - q_n}{\gamma_1}.$$

On a donc montré que, pour tout $x \in H$, on a l'estimation

$$\| (E - T_{n,0}) C^{-1} x \| \leqslant \frac{1 - q_n}{\gamma_1} \| x \|. \tag{18}$$

Comme γ_1 est la borne précise de l'opérateur autoadjoint C, γ_1 coïncide avec la valeur propre minimale de l'opérateur C. En portant dans (18) au lieu de x la fonction propre correspondant à la valeur propre minimale de l'opérateur C, on obtient dans (18) une égalité. On a obtenu, par conséquent, l'estimation $\| (E - T_{n,0}) C^{-1} \| = (1 - q_n)/\gamma_1$. Le lemme est démontré.

5. Construction de la suite optimale des paramètres d'itération *).

5.1. C a s d e $n = 2^p$. L'ordre de mise en œuvre des paramètres d'itération τ_k dans la méthode de Tchébychev influe fortement sur la convergence de la méthode. Aussi s'élève-t-il un problème de construction de meilleure suite de paramètres d'itération qui assurerait

*) Le procédé de mise en ordre des paramètres d'itération voir dans E. C. Николаев, A. A. Самарский (ЖВМ и МФ, 12, № 4, 1972), où il est donné pour tout n et [8] pour $n = 2^p$.

l'influence minimale de l'erreur de calcul de la méthode. Vu que la suite des paramètres est fonction de la mise en ordre de l'ensemble $\mathfrak{M}_n$, il faut établir dans l'ensemble $\mathfrak{M}_n$ un ordre optimal.

Donnons la solution de ce problème. Supposons d'abord que le nombre d'itérations est une puissance de 2: $n = 2^p$. Désignons par θ_m l'ensemble composé de m nombres entiers:

$$\theta_m = \{\theta_1^{(m)}, \theta_2^{(m)}, \ldots, \theta_m^{(m)}\}.$$

Sur la base de l'ensemble $\theta_1 = \{1\}$, construisons l'ensemble θ_{2^p} en appliquant la règle suivante. Soit construit l'ensemble θ_m. Alors l'ensemble θ_{2m} sera déterminé suivant les formules

$$\theta_{2m} = \{\theta_{2i}^{(2m)} = 4m - \theta_i^{(m)},\ \theta_{2i-1}^{(2m)} = \theta_i^{(m)},\ i = 1, 2, \ldots, m\},$$
$$m = 1, 2, 4, \ldots, 2^{p-1}. \quad (19)$$

On se convainc sans peine que l'ensemble θ_{2^k} est composé de nombres impairs de 1 à $2^{k+1} - 1$.

En utilisant l'ensemble construit θ_{2^p}, ordonnons l'ensemble $\mathfrak{M}_{2^p}$ de la façon suivante:

$$\mathfrak{M}_n^* = \left\{-\cos\beta_i,\ \beta_i = \frac{\pi}{2n}\theta_i^{(n)},\ i = 1, 2, \ldots, n\right\}. \quad n = 2^p. \quad (20)$$

C'est précisément l'ensemble $\mathfrak{M}_n$ ordonné cherché pour le cas où $n = 2^p$. Pour la suite des paramètres d'itération correspondant à cet ordre on a démontré l'estimation

$$\sum_{j=1}^{n} \tau_j \| T_{n,j} \| \leqslant \frac{1-q_n}{\gamma_1}.$$

En comparant cette estimation à l'estimation du lemme 1, on se convainc que l'ensemble ordonné $\mathfrak{M}_n^*$ garantit en fait l'influence minimale de l'erreur de calcul sur la convergence de la méthode de Tchébychev.

Donnons quelques exemples de construction de l'ensemble θ_n.

1) $n = 8$.

$\theta_1 = \{1\}$, $\theta_2 = \{1, 3\}$, $\theta_4 = \{1, 7, 3, 5\}$.

$\theta_8 = \{1, 15, 7, 9, 3, 13, 5, 11\}$.

L'ensemble θ_8 est construit. Suivant la formule (20) est mis en ordre l'ensemble $\mathfrak{M}_8^*$.

2) $n = 16$.

En utilisant l'ensemble θ_8 trouvé plus haut, construisons suivant les formules (19) l'ensemble θ_{16}:

$\theta_{16} = \{1, 31, 15, 17, 7, 25, 9, 23, 3, 29, 13, 19, 5, 27, 11, 21\}$.

3) $n = 32$.

$\theta_{32} = \{1, 63, 31, 33, 15, 49, 17, 47, 7, 57, 25, 39, 9, 55, 23, 41, 3, 61, 29, 35, 13, 51, 19, 45, 5, 59, 27, 37, 11, 53, 21, 43\}$.

A partir des formules (19) s'ensuit la règle simple de passage de l'ensemble θ_m à l'ensemble θ_{2m}: $\theta_{2i-1}^{(2m)} = \theta_i^{(m)}$ et la somme de deux nombres voisins vaut $4m$:

$$\theta_{2i-1}^{(2m)} + \theta_{2i}^{(2m)} = 4m, \quad i = 1, 2, \ldots, m.$$

Une règle de passage analogue s'applique également dans le cas général l'étude duquel on aborde.

5.2. Cas général. Supposons que le nombre d'itérations n soit un nombre entier quelconque. Décrivons le procédé de construction de l'ensemble θ_n. Les étapes élémentaires de ce procédé sont les passages de l'ensemble θ_m à l'ensemble θ_{2m} et de l'ensemble θ_{2m} à l'ensemble θ_{2m+1}, où m est un nombre entier quelconque.

Formulons les règles de passage d'un ensemble à l'autre.

1) Le passage de θ_{2m} à θ_{2m+1} consiste en une adjonction aux éléments de l'ensemble θ_{2m} d'un nombre impair $2m + 1$.

2) Le passage de θ_m à θ_{2m} s'effectue de la façon suivante. Si ce transfert est suivi du passage de θ_{2m} à θ_{4m} ou si le passage de θ_m à θ_{2m} est la dernière opération dans le procédé de construction de l'ensemble θ_n, on utilise les formules indiquées plus haut:

$$\theta_{2i-1}^{(2m)} = \theta_i^{(m)}, \quad \theta_{2i-1}^{(2m)} + \theta_{2i}^{(2m)} = 4m, \quad i = 1, 2, \ldots, m. \tag{21}$$

Mais si le passage de θ_m à θ_{2m} est suivi du transfert de θ_{2m} à θ_{2m+1}, on recourt aux formules

$$\theta_{2i-1}^{(2m)} = \theta_i^{(m)}, \quad \theta_{2i-1}^{(2m)} + \theta_{2i}^{(2m)} = 4m + 2, \quad i = 1, 2, \ldots, m. \tag{22}$$

En utilisant ces règles et en alternant convenablement les passages de l'ensemble à nombre pair d'éléments aux ensembles à nombre impair d'éléments et de l'ensemble à m éléments à l'ensemble à $2m$ éléments, on peut, sur la base de $\theta_1 = \{1\}$, construire l'ensemble θ_n pour tout n.

Donnons quelques exemples.

1) $n = 15$. Dans ce cas le transfert de θ_1 à θ_n s'effectue suivant la chaîne suivante:

$$\theta_1 \to \theta_2 \to \theta_3 \to \theta_6 \to \theta_7 \to \theta_{14} \to \theta_{15}.$$

Selon les règles exposées, les transferts de θ_1 à θ_2, de θ_3 à θ_6 et de θ_7 à θ_{14} s'effectuent suivant les formules (22), tandis que dans le passage de θ_2 à θ_3, de θ_6 à θ_7 et de θ_{14} à θ_{15} il faut ajouter un nombre impair correspondant à l'ensemble initial. Cela donne

$\theta_1 = \{1\}$, $\theta_2 = \{1, 5\}$, $\theta_3 = \{1, 5, 3\}$,
$\theta_6 = \{1, 13, 5, 9, 3, 11\}$, $\theta_7 = \{1, 13, 5, 9, 3, 11, 7\}$,
$\theta_{14} = \{1, 29, 13, 17, 5, 25, 9, 21, 3, 27, 11, 19, 7, 23\}$,
$\theta_{15} = \{1, 29, 13, 17, 5, 25, 9, 21, 3, 27, 11, 19, 7, 23, 15\}$.

L'ensemble $\mathfrak{M}_{15}^*$ est mis en ordre suivant la formule (20).

2) $n = 25$. A ce cas correspond la chaîne

$$\theta_1 \to \theta_2 \to \theta_3 \to \theta_6 \to \theta_{12} \to \theta_{24} \to \theta_{25},$$

les passages de θ_1 à θ_2 et de θ_{12} à θ_{24} se réalisant suivant les formules (22), tandis que les passages de θ_3 à θ_6 et de θ_6 à θ_{12} le sont suivant les formules (21), quant aux passages de θ_2 à θ_3 et de θ_{24} à θ_{25}, ils s'effectuent avec addition d'un nombre impair. Il vient

$$\theta_1 = \{1\},\ \theta_2 = \{1, 5\},\ \theta_3 = \{1, 5, 3\},\ \theta_6 = \{1, 11, 5, 7, 3, 9\},$$
$$\theta_{12} = \{1, 23, 11, 13, 5, 19, 7, 17, 3, 21, 9, 15\},$$
$$\theta_{24} = \{1, 49, 23, 27, 11, 39, 13, 37, 5, 45, 19, 31, 7, 43, 17, 33, 3, 47, 21, 29, 9, 41, 15, 35\},$$
$$\theta_{25} = \{1, 49, 23, 27, 11, 39, 13, 37, 5, 45, 19, 31, 7, 43, 17, 33, 3, 47, 21, 29, 9, 41, 15, 35, 25\}.$$

La procédure, exposée plus haut, de construction de l'ensemble θ_n pour un n quelconque peut être formalisée. A cette fin représentons n sous forme d'un développement en puissance de 2 à indices k_j entiers:

$$n = 2^{k_1} + 2^{k_2} + \ldots + 2^{k_s}, \quad k_j \leqslant k_{j-1} - 1, \quad j = 2, 3, \ldots, s.$$

Formons les quantités suivantes

$$n_j = \sum_{i=1}^{j} 2^{k_i - k_j}, \quad j = 1, 2, \ldots, s,$$

et posons $n_{s+1} = 2n + 1$. Suivant les formules (23), construisons l'ensemble θ_{n_j}:

$$\theta_{n_j} = \{\theta_i^{(n_j)} = \theta_i^{(n_j-1)},\ \theta_{n_j}^{(n_j)} = n_j,\ i = 1, 2, \ldots, n_j - 1\}, \tag{23}$$

pour $j = 1$ choisissons $\theta_1 = \{1\}$. Ensuite, suivant la formule (24), on construit l'ensemble

$$\theta_{2m} = \{\theta_{2i}^{(2m)} = 4m - \theta_i^{(m)},\ \theta_{2i-1}^{(2m)} = \theta_i^{(m)},\ i = 1, 2, \ldots, m\} \tag{24}$$

pour $m = n_j, 2n_j, 4n_j, \ldots, [(n_{j+1} - 1)/4]$, où $[a]$ est la partie entière de a. Si $[(n_{j+1} - 1)/4] < n_j$, les calculs suivant la formule (24) ne s'effectuent pas et l'on passe à l'étape suivante. Si $j = s$, l'ensemble cherché θ_n est alors construit. Au cas contraire on pose $m = (n_{j+1} - 1)/2$ et l'on construit l'ensemble

$$\theta_{2m} = \{\theta_{2i}^{(2m)} = 4m + 2 - \theta_i^{(m)},\ \theta_{2i-1}^{(2m)} = \theta_i^{(m)},\ i = 1, 2, \ldots, m\}. \tag{25}$$

Ensuite, on augmente j d'une unité et le processus se répète en commençant par la formule (23). On obtient finalement l'ensemble θ_n. L'ensemble $\mathfrak{M}_n^*$ est ordonné suivant la formule (20).

Pour le cas de $n = 2^p$ l'algorithme (23)-(25) se simplifie et passe à l'algorithme décrit par la formule (19). En effet, pour $n = 2^p$ on obtient $s = 1$, $k_1 = p$, $n_1 = 1$, $n_{s+1} = 2^{p+1} - 1$. Par conséquent, dans l'algorithme (23)-(25) j prend la seule valeur égale à l'unité et les calculs s'effectuent suivant la formule (24) pour $m = 1, 2, 4, \ldots, 2^{p-1}$.

Illustrons la qualité de la mise en ordre de l'ensemble $\mathfrak{M}_n^*$ construit ici par l'exemple étudié au point 4, § 2. Le nombre d'itérations donné n variait de 16 à 512 avec le pas 8. Pour chaque n, la précision réelle atteinte après l'exécution de n itérations ne dépassait pas la précision théorique q_n ($q_{512} = 1.23 \cdot 10^{-7}$), le processus étant en continuité de service (voir tabl. 5).

§ 3. Méthode itérative simple

1. Choix du paramètre d'itération. Au § 2 a été résolu le problème sur la construction du jeu optimal de paramètres d'itération τ_k pour un schéma à deux couches

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}} + Ay_k = f, \quad k=0,\ 1,\ \ldots, \quad y_0 \in H$$

dans l'hypothèse que l'opérateur $DB^{-1}A$ était autoadjoint dans H et qu'étaient donnés γ_1 et γ_2, les constantes de l'équivalence énergétique des opérateurs D et $DB^{-1}A$:

$$\gamma_1 D \leqslant DB^{-1}A \leqslant \gamma_2 D, \quad \gamma_1 > 0. \tag{1}$$

Cherchons maintenant la solution de ce problème pour une limitation supplémentaire $\tau_k \equiv \tau$, c'est-à-dire dans l'hypothèse que les paramètres d'itération τ_k sont indépendants du numéro d'itération k. Ce problème apparaît lorsqu'on recherche le paramètre d'itération τ pour un schéma à deux couches

$$B\frac{y_{k+1}-y_k}{\tau} + Ay_k = f, \quad k=0,\ 1,\ \ldots \tag{2}$$

Rappelons comment est formulé le problème susmentionné: parmi les polynômes de degré n de forme $Q_n(t) = \prod_{j=1}^{n} (1 - \tau_j t)$ trouver le polynôme s'écartant le moins de zéro sur le segment $[\gamma_1, \gamma_2]$. En vertu de la limite imposée, le polynôme $P_n(t)$ prend la forme

$$P_n(t) = (1 - \tau t)^n.$$

Donc le problème posé plus haut est équivalent au problème suivant: parmi les polynômes de premier degré prenant au point $t = 0$ la valeur de l'unité trouver celui qui s'écarte le moins de zéro sur le segment $[\gamma_1, \gamma_2]$.

Ce problème est un cas particulier de celui étudié au § 2. Dans le cas donné $n = 1$ et des résultats du point 1, § 2, il s'ensuit que le polynôme cherché a la forme

$$Q_1(t) = q_1 T_1\left(\frac{1-\tau_0 t}{\rho_0}\right), \quad \tau_0 = \frac{2}{\gamma_1+\gamma_2}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{\gamma_1}{\gamma_2},$$

où

$$q_1 = \frac{2\rho_1}{1+\rho_1^2} = \rho_0, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}.$$

$T_1(x)$ est ici le polynôme de Tchébychev de première espèce. Comme $T_1(x) = x$, le polynôme $Q_1(t)$ prend la forme

$$Q_1(t) = 1 - \tau_0 t, \quad \max_{\gamma_1 \leqslant t \leqslant \gamma_2} |Q_1(t)| = q_1 = \rho_0,$$

donc

$$P_n(t) = (1 - \tau_0 t)^n.$$

La valeur optimale du paramètre τ pour le schéma (2) est donc obtenue:

$$\tau = \tau_0 = 2/(\gamma_1 + \gamma_2). \tag{3}$$

Vu que la norme de l'opérateur résolvant $T_{n,0}$ pour le schéma (2) (voir point 3, § 1) est appréciée de la façon suivante:

$$\| T_{n,0} \| \leqslant \max_{\gamma_1 \leqslant t \leqslant \gamma_2} | P_n(t) |,$$

on obtiendra pour $\tau = \tau_0$ l'estimation $\| T_{n,0} \| \leqslant \rho_0^n$. De là se déduit l'estimation pour l'erreur z_n dans H_D:

$$\| z_n \|_D \leqslant \rho_0^n \| z_0 \|_D. \tag{4}$$

La méthode itérative (2), (3) s'appelle *méthode itérative simple.*

On a donc démontré le

Théorème 2. *Soit l'opérateur $DB^{-1}A$ autoadjoint satisfaisant aux conditions* (1). *La méthode itérative simple* (2), (3) *converge dans H_D, et pour l'erreur on a l'estimation* (4). *Pour le nombre d'itérations se vérifie l'estimation $n \geqslant n_0(\varepsilon)$, où $n_0(\varepsilon) = \ln \varepsilon / \ln \rho_0$.*

Remarque. Comme pour la méthode de Tchébychev, l'estimation à priori de l'erreur de la méthode itérative simple ne peut être améliorée au cas d'un espace de dimension finie.

Comparons le nombre d'itérations de la méthode de Tchébychev avec celui de la méthode itérative simple. Le théorème 1, au cas de ξ petits, fournit l'estimation suivante pour le nombre d'itérations de la méthode de Tchébychev:

$$n \geqslant n_0(\varepsilon), \quad n_0(\varepsilon) = \frac{\ln 0{,}5\varepsilon}{\ln \rho_1} \approx \frac{1}{2\sqrt{\xi}} \ln \frac{2}{\varepsilon}.$$

Le théorème 2 nous donne l'estimation pour le nombre d'itérations de la méthode itérative simple

$$n \geqslant n_0(\varepsilon), \quad n_0(\varepsilon) = \frac{\ln \varepsilon}{\ln \rho_0} \approx \frac{1}{2\xi} \ln \frac{1}{\varepsilon}.$$

Il s'ensuit de ces estimations que pour $\xi \ll 1$ le nombre d'itérations de la méthode de Tchébychev est beaucoup inférieur à celui de la méthode itérative simple. Par exemple, pour $\xi = 0{,}01$ le nombre d'itérations de la méthode itérative simple est d'environ 10 fois plus grand que pour la méthode de Tchébychev.

2. Estimation de la norme de l'opérateur de passage. Au point 1, § 3, a été étudiée la vitesse de convergence de la méthode itérative simple, cette dernière y étant considérée comme un cas particulier de la méthode de Tchébychev. Pour des raisons méthodiques il sera utile d'étudier la convergence de la méthode itérative simple indépendamment de celle de Tchébychev.

On utilisera donc pour la recherche de la solution approchée de l'équation

$$Au = f$$

le schéma à deux couches (2)

$$B\frac{y_{k+1}-y_k}{\tau}+Ay_k=f, \quad k=0,\ 1,\ \ldots, \quad y_0\in H. \tag{5}$$

Pour l'étude de la convergence du schéma (5), passons au problème posé pour une erreur équivalente $x_k = D^{1/2}z_k$:

$$x_{k+1} = Sx_k, \quad k = 0,\ 1,\ \ldots, \quad S = E - \tau C, \tag{6}$$

où $C = D^{1/2}B^{-1}AD^{-1/2}$. En utilisant (6), on trouve l'expression explicite de x_n au moyen de x_0: $x_n = S^n x_0$, d'où s'ensuit l'estimation de la norme d'erreur z_n dans H_D

$$\| z_n \|_D = \| x_n \| \leqslant \| S^n \| \, \| x_0 \| = \| S^n \| \, \| z_0 \|_D. \tag{7}$$

Supposons que l'opérateur $DB^{-1}A$ est autoadjoint dans H et que les constantes γ_1 et γ_2 sont données dans les inégalités (1). Avec ces hypothèses l'opérateur C et, en même temps, l'opérateur S sont autoadjoints dans H, γ_1 et γ_2 étant les bornes de l'opérateur C:

$$\gamma_1 E \leqslant C \leqslant \gamma_2 E, \quad \gamma_1 > 0, \quad C = C^*. \tag{8}$$

Comme l'opérateur S est autoadjoint, on a l'égalité $\| S^n \| = \| S \|^n$. Par conséquent, de l'estimation (7) il s'ensuit que le paramètre d'itération τ doit être choisi sur la base de la condition du minimum par rapport à la norme τ de l'opérateur de passage $S = E - \tau C$.

Il y a lieu au

L e m m e 2. *Soient $S = E - \tau C$ et les conditions* (8) *remplies. La norme de l'opérateur S est minimale pour $\tau = \tau_0 = 2/(\gamma_1 + \gamma_2)$ et on a l'estimation*

$$\| S \| = \| E - \tau_0 C \| = \rho_0, \quad \rho_0 = (1 - \xi)/(1 + \xi), \quad \xi = \gamma_1/\gamma_2.$$

En effet, puisque S est autoadjoint dans H, on obtient de la définition de la norme

$$\| S \| = \sup_{x\neq 0}\frac{|(Sx,\ x)|}{(x,\ x)} = \sup_{x\neq 0}\left| 1-\tau\frac{(Cx,\ x)}{(x,\ x)}\right| = \max_{\gamma_1\leqslant t\leqslant \gamma_2} | 1-\tau t |.$$

Comme $\varphi(t) = 1 - \tau t$ est une fonction linéaire, la valeur maximale en module de $\varphi(t)$ sur le segment $[\gamma_1,\ \gamma_2]$ ne sera atteinte que sur les

bouts du segment. Les calculs directs donnent

$$\|S\| = \max(|1-\tau\gamma_1|, |1-\tau\gamma_2|) = \begin{cases} \varphi_1(\tau) = 1-\tau\gamma_1, & 0 \leqslant \tau \leqslant \tau_0, \\ \varphi_2(\tau) = \tau\gamma_2 - 1, & \tau_0 \leqslant \tau, \end{cases}$$

où τ_0 est défini dans le lemme. Comme la fonction $\varphi_1(\tau)$ décroît sur le segment $[0, \tau_0]$, tandis que $\varphi_2(\tau)$ croît pour $\tau \geqslant \tau_0$, le minimum de la norme de l'opérateur S est atteint pour $\tau = \tau_0$ et est égal à $\rho_0 = 1 - \tau_0\gamma_1 = \tau_0\gamma_2 - 1 = (1-\xi)/(1+\xi)$, $\xi = \gamma_1/\gamma_2$. Le lemme est démontré.

Il découle du lemme 2 et de l'estimation (7) que pour $\tau = \tau_0$ on a pour l'erreur du schéma itératif (5) l'estimation

$$\| z_n \|_D \leqslant \rho_0^n \| z_0 \|_D.$$

On a ainsi obtenu encore une démonstration du théorème 2 formulé plus haut sur la convergence de la méthode itérative simple. Les exemples de choix de l'opérateur D pour lequel l'opérateur $DB^{-1}A$ est autoadjoint ont été étudiés au point 3, § 2.

§ 4. Cas d'opérateur non autoadjoint. Méthode itérative simple

1. Position du problème. Aux §§ 2, 3 on a construit des méthodes itératives à deux couches pour la résolution approchée de l'équation opératorielle linéaire

$$Au = f \tag{1}$$

à opérateur A non dégénéré et donné dans l'espace hilbertien réel H. On supposait que les opérateurs A, B et D étaient tels que l'opérateur $DB^{-1}A$ était autoadjoint dans H et qu'étaient données les constantes de l'équivalence énergétique γ_1 et γ_2 des opérateurs D et $DB^{-1}A$, avec $\gamma_1 > 0$.

Ces hypothèses satisfaites, on avait résolu le problème du choix optimal des paramètres d'itération et construit les méthodes de Tchébychev et itérative simple. Au point 3, § 2, on avait analysé quelques exemples de choix de l'opérateur D et trouvé les conditions pour lesquelles l'opérateur $DB^{-1}A$ était autoadjoint pour chaque choix concret de l'opérateur D.

Si les opérateurs A et B sont donnés, il n'est apparemment pas toujours possible d'indiquer l'opérateur concret D pour lequel $DB^{-1}A$ est autoadjoint dans H. Il faut donc étudier les méthodes itératives applicables à des cas d'opérateurs non autoadjoints.

On étudie dans ce paragraphe la méthode itérative simple pour le cas d'opérateurs non autoadjoints. On examinera quelques procédés de choix du paramètre d'itération en fonction du volume de l'information à priori sur les opérateurs du schéma itératif.

Supposons donc que l'opérateur $DB^{-1}A$ est *non autoadjoint dans* H. Pour une résolution approchée de l'équation (1) prenons le schéma itératif implicite à deux couches

$$B\frac{y_{k+1}-y_k}{\tau}+Ay_k=f,\quad k=0,\ 1,\ \ldots,\quad y_0\in H. \qquad (2)$$

Pour étudier la convergence du schéma (2), passons, comme habituellement, au problème posé pour l'erreur équivalente $x_k=D^{1/2}z_k$

$$x_{k+1}=Sx_k,\quad k=0,\ 1,\ \ldots,\quad S=E-\tau C, \qquad (3)$$

où $C=D^{1/2}B^{-1}AD^{-1/2}$. En vertu des hypothèses admises plus haut, l'opérateur C est non autoadjoint dans H. De la substitution faite et de l'équation (3) il vient

$$x_n=S^nx_0,\quad \|x_n\|=\|z_n\|_D\leqslant\|S^n\|\,\|x_0\|=\|S^n\|\,\|z_0\|_D. \qquad (4)$$

Par conséquent, le paramètre d'itération τ doit être choisi sur la base de la condition du minimum par rapport à la norme τ de l'opérateur résolvant S^n.

2. Minimisation de la norme de l'opérateur de passage.

2.1. P r e m i e r c a s. Cherchons l'estimation pour la norme de l'opérateur S^n. Comme pour tout opérateur on a l'estimation $\|S^n\|\leqslant\|S\|^n$, le premier mode de choix du paramètre τ consiste dans l'obtention du paramètre τ sur la base de la condition du minimum de la norme de l'opérateur de passage S. On obtient deux types d'estimation de la norme de l'opérateur S suivant le volume de l'information à priori sur l'opérateur C.

Dans le premier cas on suppose que l'information à priori consiste en la fixation de γ_1 et γ_2 des inégalités

$$\gamma_1(x,\ x)\leqslant(Cx,\ x),\quad (Cx,\ Cx)\leqslant\gamma_2(Cx,\ x),\quad \gamma_1>0. \qquad (5)$$

Si $C=C^*$, γ_1 et γ_2 sont les bornes de l'opérateur C.

L e m m e 3. *Soient données dans les inégalités* (5) γ_1 *et* γ_2, *alors se vérifie pour la norme de l'opérateur* $S=E-\tau C$ *pour* $\tau=1/\gamma_2$ *l'estimation*

$$\|S\|\leqslant\rho,\quad \rho=\sqrt{1-\xi},\quad \xi=\gamma_1/\gamma_2.$$

En effet, profitant de (5), on obtient

$$\|Sx\|^2=\|x-\tau Cx\|^2=(x,\ x)-2\tau(Cx,\ x)+\tau^2(Cx,\ Cx)\leqslant$$
$$\leqslant(x,\ x)-2\tau(Cx,\ x)+\tau^2\gamma_2(Cx,\ x)=$$
$$=\|x\|^2-\tau(2-\tau\gamma_2)(Cx,\ x).$$

Il en suit que si la condition $\tau(2-\tau\gamma_2)>0$ est remplie, c'est-à-dire si $0<\tau<2/\gamma_2$, la norme de l'opérateur S sera inférieure à l'unité. Supposons que cette condition est satisfaite, alors, en utilisant (5), il vient

$$\|Sx\|^2\leqslant[1-\tau\gamma_1(2-\tau\gamma_2)]\,\|x\|^2$$

et, partant,

$$\| S \|^2 = \sup_{x \neq 0} \frac{\| Sx \|^2}{\| x \|^2} \leqslant 1 - \tau\gamma_1 (2 - \tau\gamma_2).$$

La fonction $\varphi(\tau) = 1 - \tau\gamma_1 (2 - \tau\gamma_2)$ a un minimum au point $\tau = 1/\gamma_2$ égal à $\varphi(1/\gamma_2) = 1 - \xi$, où $\xi = \gamma_1/\gamma_2$. Donc pour la valeur indiquée du paramètre τ de la norme de l'opérateur S se vérifie l'estimation $\| S \| \leqslant \sqrt{1 - \xi}$. Le lemme est démontré.

En portant dans (5) l'opérateur $C = D^{-1/2} (DB^{-1}A) D^{-1/2}$, on obtient pour les inégalités (5) des inégalités équivalentes suivantes:

$$\gamma_1 (Dx, x) \leqslant (DB^{-1}Ax, x),$$
$$(DB^{-1}Ax, B^{-1}Ax) \leqslant \gamma_2 (DB^{-1}Ax, x), \quad \gamma_1 > 0. \tag{6}$$

En portant dans (4) l'estimation pour la norme de l'opérateur S obtenue dans le lemme 3, il vient

$$\| z_n \|_D \leqslant \rho^n \| z_0 \|_D, \quad \rho = \sqrt{1 - \xi}. \tag{7}$$

Théorème 3. *Soient γ_1 et γ_2 les constantes des inégalités* (6). *La méthode itérative simple* (2) *pour la valeur du paramètre d'itération $\tau = 1/\gamma_2$ converge dans H_D et pour l'erreur z_n on a l'estimation* (7). *Pour le nombre d'itérations se vérifie l'estimation $n \geqslant n_0(\varepsilon)$, où* $n_0(\varepsilon) = \ln \varepsilon / \ln \rho$, $\rho = \sqrt{1 - \xi}$, $\xi = \gamma_1/\gamma_2$.

Fournissons des exemples de choix de l'opérateur D ainsi que la forme concrète des inégalités (6). On a donné au tableau 6: les

Tableau 6

A et B	D	Inégalités
1) $A = A^* > 0$,	A ou $B^*A^{-1}B$	$\gamma_1 (Bx, A^{-1}Bx) \leqslant (Bx, x)$, $(Ax, x) \leqslant \gamma_2 (Bx, x)$
$B \neq B^* > 0$	A^2 ou B^*B	$\gamma_1 (Bx, Bx) \leqslant (Ax, Bx)$, $(Ax, Ax) \leqslant \gamma_2 (Ax, Bx)$
2) $A \neq A^* > 0$,	B ou $A^*B^{-1}A$	$\gamma_1 (Bx, x) \leqslant (Ax, x)$, $(Ax, B^{-1}Ax) \leqslant \gamma_2 (Ax, x)$
$B = B^* > 0$	B^2 ou A^*A	$\gamma_1 (Bx, Bx) \leqslant (Ax, Bx)$, $(Ax, Ax) \leqslant \gamma_2 (Ax, Bx)$
3) $A \neq A^* > 0$, $B \neq B^* > 0$	A^*A ou B^*B	$\gamma_1 (Bx, Bx) \leqslant (Ax, Bx)$, $(Ax, Ax) \leqslant \gamma_2 (Ax, Bx)$
4) $A = A^*$, $B = B^*$, $AB \neq BA$	A^2 ou B^2	$\gamma_1 (Bx, Bx) \leqslant (Ax, Bx)$, $(Ax, Ax) \leqslant \gamma_2 (Ax, Bx)$

hypothèses relatives aux opérateurs A et B, l'opérateur D et la forme des inégalités (6). Pour obtenir la forme concrète des inégalités (6) on se base sur les inégalités (6) elles-mêmes ainsi que sur les inégalités qui leur sont équivalentes

$$\gamma_1 (DA^{-1}Bx, A^{-1}Bx) \leqslant (DA^{-1}Bx, x), \quad (Dx, x) \leqslant \gamma_2 (DB^{-1}Ax, x), \quad (8)$$

qu'on déduit de (6) en posant $x = A^{-1}By$.

Notons les inégalités

$$\gamma_1 (Bx, Bx) \leqslant (Ax, Bx), \quad (Ax, Ax) \leqslant \gamma_2 (Ax, Bx), \quad \gamma_1 > 0.$$

Si ces conditions sont remplies, on peut, pour les cas particuliers présentés au tableau 6, choisir en qualité d'opérateur D soit l'opérateur A^2, si $A = A^*$, soit l'opérateur A^*A. Avec un tel choix de l'opérateur D la norme d'erreur z_n dans H_D peut être calculée au cours des itérations

$$\| z_n \|_D^2 = (Dz_n, z_n) = (Az_n, Az_n) = \| r_n \|^2, \quad r_n = Ay_n - f.$$

Revenons à l'estimation de la norme de l'opérateur S. Si l'opérateur C est autoadjoint dans H, alors en vertu de (5) il est défini positif et, par conséquent, il existe une racine carrée de l'opérateur C. En posant dans la seconde des inégalités (5) $x = C^{-1/2}y$, on aboutit à ce que les inégalités (5) sont équivalentes aux inégalités

$$\gamma_1 E \leqslant C \leqslant \gamma_2 E, \quad \gamma_1 > 0.$$

Il s'ensuit du lemme 2, ces hypothèses étant admises, l'estimation suivante pour la norme de l'opérateur S: $\| S \| \leqslant \rho_0$, $\rho_0 = (1 - \xi)/(1 + \xi)$, $\xi = \gamma_1/\gamma_2$.

En confrontant cette estimation à celle obtenue dans le lemme 3, on se convainc que l'estimation du lemme 3 est grossière et ne passe pas à l'estimation du lemme 2 quand l'opérateur C est autoadjoint dans H.

2.2. S e c o n d c a s. Cherchons maintenant une autre estimation pour la norme de l'opérateur de passage S, qui se transformera en l'estimation du lemme 2 quand C est un opérateur autoadjoint dans H. A cette fin augmentons le volume de l'information à priori sur l'opérateur C en admettant que sont donnés trois nombres γ_1, γ_2 et γ_3:

$$\gamma_1 E \leqslant C \leqslant \gamma_2 E, \quad \| C_1 \| \leqslant \gamma_3, \quad \gamma_1 > 0, \quad \gamma_3 \geqslant 0, \quad (9)$$

où $C_1 = 0{,}5\,(C - C^*)$ est une partie non autoadjointe de l'opérateur C.

On a le

L e m m e 4. *Soient donnés γ_1, γ_2 et γ_3 dans les inégalités (9). Dans ce cas pour la norme de l'opérateur $S = E - \tau C$ pour $\tau =$*

$= \bar{\tau}_0 = \tau_0 (1 - \varkappa\bar{\rho}_0)$ *se vérifie l'estimation*

$$\| S \| \leqslant \bar{\rho}_0, \quad \bar{\rho}_0 = (1 - \bar{\xi})/(1 + \bar{\xi}), \tag{9'}$$

où

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \varkappa = \frac{\gamma_3}{\sqrt{\gamma_1\gamma_2 + \gamma_3^2}}, \quad \bar{\xi} = \frac{1-\varkappa}{1+\varkappa} \cdot \frac{\gamma_1}{\gamma_2}.$$

Esquissons la démonstration du lemme 4. Soit θ un nombre quelconque de l'intervalle (0, 1). Représentons l'opérateur S sous la forme suivante:

$$S = E - \tau C = [\theta E - \tau C_0] + [(1 - \theta) E - \tau C_1],$$

où $C_0 = 0{,}5\,(C + C^*)$ est la partie autoadjointe de l'opérateur C. En recourant à l'inégalité du triangle, on aboutit à l'estimation pour la norme de l'opérateur S:

$$\| S \| \leqslant \| \theta E - \tau C_0 \| + \| (1 - \theta) E - \tau C_1 \|. \tag{10}$$

Apprécions séparément la norme de chaque opérateur. A partir de (9) et de l'égalité $(C_0 x, x) = 0{,}5\,(Cx, x) + 0{,}5\,(C^* x, x) = (Cx, x)$, on obtient que γ_1 et γ_2 sont les bornes de l'opérateur autoadjoint C_0:

$$\gamma_1 E \leqslant C_0 \leqslant \gamma_2 E, \quad \gamma_1 > 0.$$

Par analogie avec le lemme 2 on obtient l'estimation suivante pour la norme de l'opérateur $\theta E - \tau C_0$:

$$\| \theta E - \tau C_0 \| \leqslant \max (| \theta - \tau\gamma_1 |, | \theta - \tau\gamma_2 |) = \begin{cases} \theta - \tau\gamma_1, & 0 \leqslant \tau \leqslant \theta\tau_0, \\ \tau\gamma_2 - \theta, & \tau \geqslant \theta\tau_0. \end{cases}$$

Apprécions la norme de l'opérateur $(1 - \theta) E - \tau C_1$. Vu que $(C_1 x, x) = 0$, on obtient pour tous les $x \in H$:

$$\| ((1 - \theta) E - \tau C_1) x \|^2 =$$
$$= (1 - \theta)^2 \| x \|^2 + \tau^2 \| C_1 x \|^2 \leqslant ((1 - \theta)^2 + \tau^2 \| C_1 \|^2) \| x \|^2.$$

De là et de (9) s'ensuit l'estimation $\| (1 - \theta) E - \tau C_1 \| \leqslant [(1 - \theta)^2 + \tau^2\gamma_3^2]^{1/2}$. En portant l'estimation obtenue dans (10), il vient

$$\| S \| \leqslant \begin{cases} \varphi_1 (\theta, \tau) = \theta - \tau\gamma_1 + \sqrt{(1 - \theta)^2 + \tau^2\gamma_3^2}, & 0 \leqslant \tau \leqslant \tau_0\theta, \\ \varphi_2 (\theta, \tau) = \tau\gamma_2 - \theta + \sqrt{(1 - \theta)^2 + \tau^2\gamma_3^2}, & \tau \geqslant \tau_0\theta. \end{cases}$$

Choisissons maintenant les paramètres τ et θ sur la base de la condition du minimum de l'estimation pour la norme de l'opérateur S. Notons que la fonction $\varphi_2 (\theta, \tau)$ croît de façon monotone en τ. Aussi pour la minimisation de la norme de l'opérateur S suffit-il d'étudier le domaine $0 \leqslant \tau \leqslant \tau_0\theta$, $0 < \theta < 1$. Dans ce domaine $\| S \| \leqslant \varphi_1 (\theta, \tau)$.

Etudions la fonction $\varphi_1 (\theta, \tau)$. Cette fonction croît de façon monotone en θ, aussi le minimum est-il atteint pour $\tau = \tau_0\theta$. Avec cette valeur du paramètre τ on aura

$$\| S \| \leqslant \varphi (\theta) = \varphi_1 (\theta, \tau_0\theta) = \theta (1 - \tau_0\gamma_1) + \sqrt{(1 - \theta)^2 + \tau_0^2\gamma_3^2\theta^2} =$$
$$= \theta\rho_0 + \sqrt{(1 - \theta)^2 + \tau_0^2\gamma_3^2\theta^2}.$$

Bref, il faut démontrer que $\min\limits_{0<\theta<1} \varphi (\theta) = \bar{\rho}_0$. Cherchons le minimum de la

fonction $\varphi(\theta)$. Faisons la substitution de variable en posant

$$\theta = (1 - x)/(1 + a^2), \quad x \in (-a^2, 1), \quad a^2 = \tau_0^2\gamma_3^2.$$

La fonction $\varphi(\theta)$ se transcrit sous la forme

$$\varphi(\theta) = \bar{\varphi}(x) = \frac{1}{\sqrt{1+a^2}}\left(\sqrt{x^2+a^2} - \frac{\rho_0}{\sqrt{1+a^2}}x\right) + \frac{\rho_0}{1+a^2}. \tag{11}$$

De là on voit qu'il suffit de trouver le minimum de la fonction

$$v(x) = \sqrt{x^2+a^2} - \rho_0 x/\sqrt{1+a^2}$$

dans le domaine $-a^2 < x < 1$. En calculant les dérivées de la fonction $v(x)$

$$v'(x) = \frac{x}{\sqrt{x^2+a^2}} - \frac{\rho_0}{\sqrt{1+a^2}}, \quad v''(x) = \frac{a^2}{(x^2+a^2)^{3/2}} > 0,$$

on trouve que l'équation $v'(x) = 0$ fournit le point du minimum de la fonction $v(x)$. En résolvant l'équation

$$\sqrt{\frac{x^2+a^2}{1+a^2}} = \frac{x}{\rho_0}, \tag{12}$$

on obtient le point cherché du minimum de la fonction $v(x)$:

$$x_0 = a\rho_0/\sqrt{1+a^2-\rho_0^2} \in (0, 1), \quad \theta_0 = (1-x_0)/(1+a^2).$$

Portant (12) dans (11), on aboutit à la valeur minimale de la fonction $\varphi(\theta)$:

$$\varphi(\theta_0) = \sqrt{\frac{x_0^2+a^2}{1+a^2}} + \rho_0\frac{1-x_0}{1+a^2} = \frac{x_0}{\rho_0} + \theta_0\rho_0. \tag{13}$$

Il ne reste qu'à exprimer x_0 et θ_0 en fonction des valeurs connues. En se servant des notations du lemme 4, on obtient

$$1-\rho_0^2 = \tau_0^2\gamma_1\gamma_2, \quad x_0 = \tau_0\gamma_3\rho_0/\sqrt{1-\rho_0^2+\tau_0^2\gamma_3^2} = \varkappa\rho_0. \tag{14}$$

A partir de (12) il vient

$$a^2 = x_0^2(1-\rho_0^2)/(\rho_0^2 - x_0^2), \quad 1 + a^2 = \rho_0^2(1 - x_0^2)/(\rho_0^2 - x_0^2).$$

Donc

$$\theta_0\rho_0 = \frac{1-x_0}{1+a^2}\rho_0 = \frac{\rho_0^2 - x_0^2}{\rho_0(1+x_0)} = \frac{\rho_0(1-\varkappa^2)}{1+\varkappa\rho_0}. \tag{15}$$

Portons (14) et (15) dans (13) :

$$\varphi(\theta_0) = \varkappa + \frac{\rho_0(1-\varkappa^2)}{1+\rho_0\varkappa} = \frac{\varkappa+\rho_0}{1+\rho_0\varkappa} = \frac{(1+\varkappa)-\xi(1-\varkappa)}{(1+\varkappa)+\xi(1-\varkappa)} = \frac{1-\bar{\xi}}{1+\bar{\xi}} = \bar{\rho}_0. \tag{16}$$

Cherchons maintenant l'expression pour le paramètre $\tau = \tau_0\theta_0$. En comparant (15) et (16), il vient

$$\theta_0\rho_0 = \bar{\rho}_0 - \varkappa. \tag{17}$$

D'autre part, de (16) on peut tirer ρ_0 en l'exprimant en fonction de $\bar{\rho}_0$ et $\varkappa$:

$$\rho_0 = (\bar{\rho}_0 - \varkappa)/(1 - \varkappa\bar{\rho}_0).$$

En portant ρ_0 dans (17), on trouve

$$\theta_0 = 1 - \varkappa\bar{\rho}_0, \quad \tau = \tau_0(1 - \varkappa\bar{\rho}_0).$$

Le lemme est démontré.

Les inégalités (9) peuvent être écrites sous la forme suivante:

$$\gamma_1(x, x) \leqslant (Cx, x) \leqslant \gamma_2(x, x), \quad (C_1x, C_1x) \leqslant \gamma_3^2(x, x), \quad \gamma_1 > 0.$$

En y portant $C = D^{-1/2}(DB^{-1}A)D^{-1/2}$ et $C_1 = 0{,}5D^{-1/2}(DB^{-1}A - (DB^{-1}A)^*)D^{-1/2}$, on obtient les inégalités

$$\gamma_1 D \leqslant DB^{-1}A \leqslant \gamma_2 D, \quad \gamma_1 > 0,$$
$$\left(D^{-1}\frac{DB^{-1}A-(DB^{-1}A)^*}{2}x,\ \frac{DB^{-1}A-(DB^{-1}A)^*}{2}x\right) \leqslant \gamma_3^2 (Dx, x). \tag{18}$$

En portant dans (4) l'estimation (9′) pour la norme de l'opérateur S, il vient

$$\| z_n \|_D \leqslant \bar{\rho}_0^n \| z_0 \|_D. \tag{19}$$

T h é o r è m e 4. *Soient* γ_1, γ_2 *et* γ_3 *les constantes dans les inégalités* (18). *La méthode itérative simple* (2) *pour la valeur du paramètre d'itération* $\tau = \bar{\tau}_0 = \tau_0(1 - \varkappa\bar{\rho}_0)$ *converge dans* H_D *et on a pour l'erreur* z_n *l'estimation* (19). *Pour le nombre d'itérations se vérifie l'estimation* $n \geqslant n_0(\varepsilon)$, *où* $n_0(\varepsilon) = \ln \varepsilon / \ln \bar{\rho}_0$,

$$\tau_0 = \frac{2}{\gamma_1+\gamma_2}, \quad \bar{\rho}_0 = \frac{1-\bar{\xi}}{1+\bar{\xi}}, \quad \bar{\xi} = \frac{1-\varkappa}{1+\varkappa}\cdot\frac{\gamma_1}{\gamma_2}, \quad \varkappa = \frac{\gamma_3}{\sqrt{\gamma_1\gamma_2+\gamma_3^2}}.$$

R e m a r q u e. Comme le nombre d'itérations se définit par la valeur de $\bar{\xi}$, qui peut être écrite sous forme

$$\bar{\xi} = \left(\sqrt{\gamma_1/\gamma_2 + (\gamma_3/\gamma_2)^2} - \gamma_3/\gamma_2\right)^2,$$

l'opérateur B doit être choisi de manière que le rapport $\xi = \gamma_1/\gamma_2$ soit maximal, et γ_3/γ_2 minimal.

Donnons des exemples de choix de l'opérateur D. Si en guise de D on choisit l'opérateur A^*A ou B^*B, les inégalités (18) pourront être écrites sous forme

$$\gamma_1(Bx, Bx) \leqslant (Ax, Bx) \leqslant \gamma_2(Bx, Bx),$$
$$\| 0{,}5(AB^{-1} - (B^*)^{-1}A^*) \| \leqslant \gamma_3. \tag{20}$$

En effet, au cas de $D = B^*B$ cette assertion est évidente, mais si $D = A^*A$, il faut procéder dans (18) à la substitution $x = A^{-1}By$ et obtenir les inégalités (20).

Si l'opérateur B est autoadjoint, défini positif et borné dans H, on peut, en qualité d'opérateur D, prendre l'opérateur B ou $A^*B^{-1}A$. Dans ce cas les inégalités (18) sont équivalentes aux inégalités suivantes:

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \quad \gamma_1 > 0,$$
$$(B^{-1}A_1x, A_1x) \leqslant \gamma_3^2(Bx, x), \quad A_1 = 0{,}5(A - A^*). \tag{21}$$

En effet, pour $D = B$ les inégalités (18) et (21) coïncident, tandis que pour $D = A^*B^{-1}A$ les inégalités (21) découlent des inégalités (18) après substitution de $x = A^{-1}By$ dans (18).

3. Minimisation de la norme de l'opérateur résolvant.

3.1. P r e m i e r c a s. Au point 2, § 4, on a obtenu les estimations pour les normes de l'opérateur S^n basées sur l'inégalité $\| S^n \| \leqslant \| S \|^n$. Voyons

maintenant un autre procédé permettant d'obtenir l'estimation pour $\| S^n \|$. Ce procédé se fonde sur l'estimation du rayon numérique de l'opérateur.

Rappelons (voir § 1, ch. V) que par le *rayon numérique de l'opérateur* T, agissant dans l'espace hilbertien complexe $\widetilde{H}$, on appelle la quantité

$$\rho(T) = \sup_{\|z\|=1} |(Tz, z)|, \quad z \in \widetilde{H}.$$

Pour l'opérateur linéaire borné T le rayon numérique vérifie les inégalités

$$\mu(T) \| T \| \leqslant \rho(T) \leqslant \| T \|, \quad \rho(T^n) \leqslant [\rho(T)]^n, \tag{22}$$

où n est un nombre naturel, tandis que $\mu(T) \geqslant 1/2$.

En profitant de la notion de rayon numérique de l'opérateur, on aboutit à deux estimations pour la norme de l'opérateur S^n suivant le type de l'information à priori sur l'opérateur C.

Voyons le cas où l'information à priori est donnée sous forme des constantes γ_1, γ_2 et γ_3:

$$\gamma_1 E \leqslant C \leqslant \gamma_2 E, \quad \| C_1 x \| \leqslant \gamma_3 \| x \|, \quad \gamma_1 > 0, \quad x \in H. \tag{23}$$

L'espace complexe $\widetilde{H}$ sera défini de la façon suivante: il est composé d'éléments de la forme $z = x + iy$, où x, $y \in H$. Le produit scalaire dans $\widetilde{H}$ se définit par la formule

$$(z, w) = (x, u) + i(y, u) - i(x, v) + (y, v),$$
$$z = x + iy, \quad w = u + iv.$$

L'opérateur linéaire C donné sur H est défini sur $\widetilde{H}$ de la manière suivante: $Cz = Cx + iCy$.

En vertu des propriétés (22), pour tout nombre entier n se vérifie l'estimation

$$\| S^n \| \leqslant \frac{1}{\mu(S^n)} \rho(S^n) \leqslant 2 [\rho(S)]^n,$$

aussi suffit-il d'apprécier le rayon numérique de l'opérateur S.

On a le

Lemme 5. *Soient* γ_1, γ_2 *et* γ_3 *données dans les inégalités* (23). *Alors pour la norme de l'opérateur* $S = E - \tau C$ *dans* H *pour* $\tau = \min(\tau_0, \varkappa\tau_0)$ *se vérifie l'estimation*

$$\| S^n \| \leqslant 2\rho^n,$$

où

$$\rho^2 = \begin{cases} 1 - \varkappa(1 - \rho_0), & 0 < \varkappa \leqslant 1, \\ 1 - (2 - 1/\varkappa)(1 - \rho_0), & \varkappa \geqslant 1, \end{cases} \quad \varkappa = \frac{\gamma_1(\gamma_1 + \gamma_2)}{2(\gamma_1^2 + \gamma_3^2)},$$

$$\tau_0 = 2/(\gamma_1 + \gamma_2), \quad \rho_0 = (1 - \xi)/(1 + \xi), \quad \xi = \gamma_1/\gamma_2.$$

Pour démontrer le lemme, représentons l'opérateur C sous forme de somme $C = C_0 + C_1$, $C_0 = 0{,}5\,(C + C^*)$, $C_1 = 0{,}5\,(C - C^*)$. Apprécions le rayon numérique de l'opérateur $S = E - \tau C$. On obtient pour tout $z \in \widetilde{H}$

$$(Sz, z) = (z, z) - \tau(C_0 z, z) - \tau(C_1 z, z).$$

En vertu du fait que l'opérateur C_0 est autoadjoint, le produit scalaire $(C_0 z, z)$ est un nombre réel. Puisque $C_1 = -C_1^*$, $(C_1 z, z)$ est un nombre imaginaire. Donc

$$|(Sz, z)|^2 = [(z, z) - \tau(C_0 z, z)]^2 + \tau^2 |(C_1 z, z)|^2. \tag{24}$$

A partir des inégalités (23) on tire

$$\gamma_1(z, z) \leqslant (C_0 z, z) \leqslant \gamma_2(z, z), \quad \| C_1 z \| \leqslant \gamma_3 \| z \|. \tag{25}$$

Soit $\|z\| = 1$. De (25) il vient

$$|(z, z) - \tau (C_0 z, z)| \leqslant \max (|1-\tau\gamma_1|, |1-\tau\gamma_2|) = \begin{cases} 1-\tau\gamma_1, & 0 \leqslant \tau \leqslant \tau_0, \\ \tau\gamma_2 - 1, & \tau \geqslant \tau_0, \end{cases}$$

$$|(C_1 z, z)| \leqslant \|C_1 z\| \, \|z\| \leqslant \gamma_3.$$

En portant ces estimations dans (24), il vient

$$\rho^2 (S) = \sup_{\|z\|=1} |(Sz, z)|^2 \leqslant \begin{cases} \varphi_1 (\tau) = (1-\tau\gamma_1)^2 + \tau^2\gamma_3^2, & 0 \leqslant \tau \leqslant \tau_0, \\ \varphi_2 (\tau) = (1-\tau\gamma_2)^2 + \tau^2\gamma_3^2, & \tau \geqslant \tau_0. \end{cases}$$

Choisissons le paramètre τ sur la base de la condition du minimum d'appréciation du rayon numérique de l'opérateur S. Vu que la fonction $\varphi_2(\tau)$ croît en τ pour $\tau \geqslant \tau_0$:

$$\varphi_2' (\tau) = 2 [\tau (\gamma_2^2 + \gamma_3^2) - \gamma_2] \geqslant 2 \frac{\gamma_2 (\gamma_2 - \gamma_1) + 2\gamma_3^2}{\gamma_1 + \gamma_2} > 0,$$

le minimum $\rho(S)$ en τ doit donc être cherché dans le domaine $\tau \leqslant \tau_0$, où pour $\rho(S)$ est vérifiée l'estimation $\rho^2(S) \leqslant \varphi_1(\tau)$.

Etudions la fonction $\varphi_1(\tau)$. Comme

$$\varphi_1'' (\tau) = 2 (\gamma_1^2 + \gamma_3^2) > 0,$$

en égalant la dérivée

$$\varphi_1' (\tau) = 2 [\tau (\gamma_1^2 + \gamma_3^2) - \gamma_1]$$

à zéro, on trouve le point extrémal de la fonction $\varphi_1(\tau)$

$$\tau = \tau_1 = \frac{\gamma_1}{\gamma_1^2 + \gamma_3^2} = \tau_0 \varkappa.$$

Pour $\tau \leqslant \tau_1$ la fonction $\varphi_1(\tau)$ décroît, tandis que pour $\tau \geqslant \tau_1$ elle s'accroît. Aussi la valeur minimale de $\varphi_1(\tau)$ est-elle atteinte au point $\tau = \tau_1$, si $\tau_1 \leqslant \tau_0$, et au point $\tau = \tau_0$, si $\tau_1 \geqslant \tau_0$. On a donc trouvé la valeur optimale du paramètre τ, $\tau = \min(\tau_0, \tau_0\varkappa)$. De plus

$$\min_\tau \rho^2 (S) \leqslant \begin{cases} \varphi_1 (\tau_0), & \varkappa \geqslant 1, \\ \varphi_1 (\tau_1), & 0 \leqslant \varkappa \leqslant 1. \end{cases}$$

Calculons $\varphi_1(\tau_0)$ et $\varphi_1(\tau_1)$. A partir de la définition de $\varkappa$ et de l'égalité $1 - \tau_0\gamma_1 = \rho_0$, on obtient

$$\varkappa = \frac{\tau_0\gamma_1}{\tau_0^2\gamma_1^2 + \tau_0^2\gamma_3^2}, \quad \tau_0^2\gamma_3^2 = \frac{\tau_0\gamma_1}{\varkappa} - 1 - \tau_0^2\gamma_1^2 = \frac{1-\rho_0}{\varkappa} - (1-\rho_0)^2.$$

Ensuite,

$$\varphi_1 (\tau_0) = (1 - \tau_0\gamma_1)^2 + \tau_0^2\gamma_3^2 = \rho_0^2 + (1-\rho_0)/\varkappa - (1-\rho_0)^2 = \\ = 1 - (2 - 1/\varkappa)(1 - \rho_0),$$

$$\varphi_1 (\tau_1) = (1 - \tau_1\gamma_1)^2 + \tau_1^2\gamma_3^2 = 1 - 2\tau_1\gamma_1 + \tau_1^2 (\gamma_1^2 + \gamma_3^2) = \\ = 1 - \tau_1\gamma_1 = 1 - \varkappa\tau_0\gamma_1 = 1 - \varkappa(1-\rho_0).$$

Bref, le rayon numérique est apprécié. L'estimation du lemme s'ensuit de l'inégalité $\|S^n\| \leqslant 2 [\rho(S)]^n$. Le lemme est démontré.

En recourant au lemme 5 on obtient l'estimation pour la norme d'erreur z_n:

$$\|z_n\|_D \leqslant 2\rho^n \|z_0\|_D. \tag{26}$$

Théorème 5. *Soient* γ_1, γ_2 *et* γ_3 *les constantes dans les inégalités* (18). *La méthode itérative simple* (2) *pour la valeur du paramètre d'itération* $\tau = \min(\tau_0, \varkappa\tau_0)$ *converge dans* H_D, *et pour l'erreur* z_n *on a l'estimation* (26). *Pour le nombre d'itérations se vérifie l'estimation* $n \geqslant n_0(\varepsilon)$, *où* $n_0(\varepsilon) = \ln(0{,}5\varepsilon)/\ln\rho$, $\varkappa$ *et* ρ *étant définis dans le lemme* 5.

Les exemples de choix de l'opérateur D et l'aspect concret des inégalités (18) sont donnés au point 2.2.

3.2. Second cas. En recourant à la notion de rayon numérique de l'opérateur, on obtient encore une estimation pour la norme de l'opérateur S^n. Admettons que l'information à priori est donnée sous forme de constantes γ_1, γ_2 et γ_3 dans les inégalités

$$\gamma_1 E \leqslant C \leqslant \gamma_2 E, \quad (C_1 x, C_1 x) \leqslant \gamma_3 (Cx, x), \quad \gamma_1 > 0. \tag{27}$$

On a le

Lemme 6. *Soient γ_1, γ_2 et γ_3 données dans les inégalités* (27). *Pour la norme de l'opérateur $S = E - \tau C$ dans H pour $\tau = \min(\tau_0^*, \varkappa\tau_0^*)$ se vérifie alors l'estimation*

$$\| S^n \| \leqslant 2\rho^n,$$

où

$$\rho^2 = \begin{cases} 1-(2\varkappa-1)\dfrac{1-\rho_0}{1+\rho_0}, & \dfrac{1}{2} \leqslant \varkappa \leqslant 1, \\ 1-\left(2-\dfrac{1}{\varkappa}\right)^2 \dfrac{1-\rho_0}{1+\rho_0}, & \varkappa \geqslant 1, \end{cases} \qquad \varkappa = \frac{\gamma_1+\gamma_2+\gamma_3}{2(\gamma_1+\gamma_3)} .$$

$$\tau_0^* = 2/(\gamma_1+\gamma_2+\gamma_3), \quad \rho_0 = (1-\xi)/(1+\xi), \quad \xi = \gamma_1/\gamma_2.$$

Et de fait, en représentant l'opérateur C sous forme de $C = C_0 + C_1$, où $C_0 = 0{,}5\,(C + C^*)$ et $C_1 = 0{,}5\,(C - C^*)$, il vient

$$|(Sz, z)|^2 = [(z, z) - \tau (C_0 z, z)]^2 + \tau^2 |(C_1 z, z)|^2.$$

A partir de l'inégalité de Cauchy-Bouniakovski et des conditions du lemme on obtient

$$|(C_1 z, z)|^2 \leqslant (C_1 z, C_1 z)(z, z) \leqslant \gamma_3 (C_0 z, z)(z, z).$$

Vu que pour tout $z \in \widehat{H}$ on a les inégalités

$$\gamma_1 (z, z) \leqslant (C_0 z, z) \leqslant \gamma_2 (z, z), \quad \gamma_1 > 0,$$

à partir des trois relations précédentes on peut donc déduire l'estimation suivante pour le rayon numérique de l'opérateur S:

$$\rho^2(S) \leqslant \max_{\gamma_1 \leqslant t \leqslant \gamma_2} \varphi(t), \text{ où } \varphi(t) = (1-\tau t)^2 + \tau^2 \gamma_3 t.$$

Etudions la fonction $\varphi(t)$. Cette fonction ne peut prendre une valeur maximale qu'aux bouts du segment $[\gamma_1, \gamma_2]$. Donc

$$\rho^2(S) \leqslant \begin{cases} \varphi_1(\tau) = (1-\tau\gamma_1)^2 + \tau^2\gamma_1\gamma_3, & 0 \leqslant \tau \leqslant \tau_0^*, \\ \varphi_2(\tau) = (1-\tau\gamma_2)^2 + \tau^2\gamma_2\gamma_3, & \tau \geqslant \tau_0^*. \end{cases}$$

Choisissons le paramètre τ à partir de la condition du minimum de l'estimation de $\rho(S)$. Comme la fonction $\varphi_2(\tau)$ croît en τ pour $\tau \geqslant \tau_0^*$:

$$\varphi_2'(\tau) = 2\gamma_2 [\tau(\gamma_2+\gamma_3) - 1] \geqslant 2\gamma_2 \frac{\gamma_2 - \gamma_1 + \gamma_3}{\gamma_2 + \gamma_1 + \gamma_3} > 0,$$

le minimum $\rho(S)$ doit être recherché dans le domaine $\tau \leqslant \tau_0^*$, où pour $\rho(S)$ se vérifie l'estimation $\rho^2(S) \leqslant \varphi_1(\tau)$.

La fonction $\varphi_1(\tau)$ pour $\tau = \tau_1 = 1/(\gamma_1 + \gamma_3) = \varkappa\tau_0^*$ atteint la valeur minimale, mais pour $\tau \leqslant \tau_1$ la fonction $\varphi_1(\tau)$ décroît, tandis que pour $\tau \geqslant \tau_1$ elle croît. Aussi la valeur minimale de $\varphi_1(\tau)$ sur le segment $[0, \tau_0^*]$ est-elle atteinte au point $\tau = \tau_1$, si $\tau_1 \leqslant \tau_0^*$, et au point $\tau = \tau_0^*$, si $\tau_1 \geqslant \tau_0^*$.

Finalement on a trouvé la valeur optimale du paramètre τ:

$$\tau = \min(\tau_0^*, \varkappa\tau_0^*).$$

De plus,

$$\min_{\tau} \rho^2(S) \leqslant \begin{cases} \varphi_1(\tau_0^*), & \varkappa \geqslant 1, \\ \varphi_1(\tau_1), & \varkappa \leqslant 1. \end{cases}$$

Calculons $\varphi_1(\tau_0^*)$ et $\varphi_1(\tau_1)$. Des calculs simples fournissent $\tau_0^* = (2 - 1/\varkappa)/\gamma_2$, $\gamma_3 = 1/(\varkappa\tau_0^*) - \gamma_1$. En utilisant ces relations, on obtient

$$\varphi_1(\tau_0^*) = (1 - \tau_0^*\gamma_1)^2 + (\tau_0^*)^2\gamma_1\gamma_3 = 1 - 2\tau_0^*\gamma_1 + \tau_0^*\gamma_1/\varkappa =$$
$$= 1 - (2 - 1/\varkappa)^2\gamma_1/\gamma_2 = 1 - (2 - 1/\varkappa)^2 (1 - \rho_0)/(1 + \rho_0).$$

Ensuite,

$$\varphi_1(\tau_1) = (1 - \tau_0^*\varkappa\gamma_1)^2 + (\tau_0^*)^2\varkappa^2\gamma_1\gamma_3 = 1 - \tau_0^*\varkappa\gamma_1 =$$
$$= 1 - (2\varkappa - 1)\gamma_1/\gamma_2 = 1 - (2\varkappa - 1)(1 - \rho_0)/(1 + \rho_0).$$

On a ainsi obtenu l'estimation du rayon numérique. L'estimation du lemme est tirée de l'inégalité $\| S^n \| \leqslant 2 [\rho(S)]^n$. Le lemme est démontré.

En portant dans les inégalités (27) $C = D^{-1/2}(DB^{-1}A)D^{-1/2}$ et $C_1 = 0{,}5D^{-1/2}(DB^{-1}A - (DB^{-1}A)^*)D^{-1/2}$, on obtient les inégalités

$$\gamma_1 D \leqslant DB^{-1}A \leqslant \gamma_2 D, \quad \gamma_1 > 0,$$
$$\left(D^{-1}\frac{DB^{-1}A-(DB^{-1}A)^*}{2}x, \ \frac{DB^{-1}A-(DB^{-1}A)^*}{2}x\right) \leqslant \gamma_3 (DB^{-1}Ax, x). \tag{28}$$

Théorème 6. *Soient γ_1, γ_2 et γ_3 les constantes dans les inégalités* (28). *La méthode itérative simple* (2) *pour la valeur du paramètre d'itération* $\tau = \min(\tau_0^*, \varkappa\tau_0^*)$ *converge dans H_D, et pour l'erreur z_n se vérifie l'estimation* (26). *Pour le nombre d'itérations on a l'estimation $n \geqslant n_0(\varepsilon)$, où*

$$n_0(\varepsilon) = \ln 0{,}5\varepsilon/\ln\rho,$$

$\varkappa$, ρ *et* τ_0^* *étant définis dans le lemme* 6.

Fournissons l'aspect des inégalités (28) pour quelques exemples de choix de l'opérateur D. Si en guise d'opérateur D on prend l'opérateur A^*A ou B^*B, les inégalités (28) peuvent être écrites sous la forme suivante:

$$\gamma_1(Bx, Bx) \leqslant (Ax, Bx) \leqslant \gamma_2(Bx, Bx), \quad \gamma_1 > 0,$$
$$\| 0{,}5(AB^{-1} - (B^*)^{-1}A^*)x \|^2 \leqslant \gamma_3(A^*x, B^{-1}x). \tag{29}$$

En effet, les inégalités (29) dérivent directement de (28) après substitution de $D = B^*B$ dans (29). Au cas de $D = A^*A$ il suffit dans (28) de procéder à la substitution $x = A^{-1}By$.

Si l'opérateur B est autoadjoint défini positif dans H et borné, on peut prendre en qualité d'opérateur D les opérateurs B ou $A^*B^{-1}A$. Dans ce cas les inégalités (28) auront la forme

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \quad \gamma_1 > 0,$$
$$(B^{-1}A_1x, A_1x) \leqslant \gamma_3(Ax, x), \quad A_1 = 0{,}5(A - A^*). \tag{30}$$

Notons qu'au cas de $D = A^*B^{-1}A$ les inégalités (30) se déduisent de (28) après substitution mentionnée plus haut.

4. Méthode de symétrisation de l'équation. Dans la résolution de l'équation $Au = f$ à opérateur A non autoadjoint on recourt à un procédé bien connu de *symétrisation de l'équation*. Au lieu de l'équation primitive on étudie l'équation symétrisée

$$\tilde{A}u = \tilde{f}, \quad \tilde{A} = A^*A, \quad \tilde{f} = A^*f, \tag{31}$$

qu'on obtient à partir de l'équation de départ en la multipliant à gauche par l'opérateur adjoint à A. En algèbre cette transformation de l'équation porte le nom de *première transformation de Gauss*.

Pour obtenir la solution approchée de l'équation (31), prenons le schéma implicite à deux couches

$$\tilde{B}\frac{y_{k+1}-y_k}{\tau_{k+1}}+\tilde{A}y_k=\tilde{f}, \quad k=0,\ 1,\ \ldots, \quad y_0\in H, \tag{32}$$

avec l'opérateur $\tilde{B}$ autoadjoint et défini positif. En guise d'opérateur D choisissons les opérateurs $\tilde{B}$ ou $\tilde{A} = A^*A$. Dans ce cas l'opérateur $D\tilde{B}^{-1}\tilde{A}$ est autoadjoint dans H et, par suite, les paramètres d'itération τ_k peuvent être choisis suivant les formules de la méthode de Tchébychev étudiée au § 2. L'information à priori de cette méthode pour le cas des opérateurs D mentionnés prend l'aspect de constantes de l'équivalence énergétique des opérateurs $\tilde{B}$ et $\tilde{A} = A^*A$

$$\gamma_1\tilde{B}\leqslant A^*A\leqslant\gamma_2\tilde{B}, \quad \gamma_1>0.$$

L'estimation de la vitesse de convergence de la méthode de Tchébychev (32) et les formules des paramètres d'itération sont données dans le théorème 1.

§ 5. Exemples d'application des méthodes itératives

1. Problème de différences de Dirichlet pour l'équation de Poisson dans un rectangle. Afin d'illustrer l'application des méthodes itératives à deux couches construites dans ce chapitre, étudions comment se résout le problème de différences de Dirichlet pour des équations elliptiques linéaires de second ordre. Le problème de différences sera traité comme une équation opératorielle

$$Au = f \tag{1}$$

dans un espace de dimension finie des fonctions de mailles. On examinera les méthodes de Tchébychev explicite et implicite ainsi que la méthode itérative simple.

Commençons l'étude des exemples par le *problème de Dirichlet pour le problème de Poisson dans un rectangle.* Supposons que dans le rectangle $\overline{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha,\ \alpha = 1,\ 2\}$ à frontière Γ il s'agit de trouver la solution de l'équation de Poisson

$$Lu=\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}=-f(x), \quad x=(x_1,\ x_2)\in G, \tag{2}$$

qui prend à la frontière Γ les valeurs données

$$u(x) = g(x), \quad x \in \Gamma. \tag{3}$$

Le *problème de différences de Dirichlet* correspondant à (2), (3) sur le maillage

$$\overline{\omega}=\{x_{ij}=(ih_1,\ jh_2)\in\overline{G}, \quad 0\leqslant i\leqslant N_1, \quad 0\leqslant j\leqslant N_2, \quad h_\alpha=l_\alpha/N_\alpha,\ \alpha=1,\ 2\}$$

prend la forme

$$\Lambda y = \sum_{\alpha=1}^{2} y_{\bar{x}_\alpha x_\alpha} = -\varphi(x), \quad x \in \omega, \quad y(x) = g(x), \quad x \in \gamma, \tag{4}$$

où $\gamma = \{x_{ij} \in \Gamma\}$ est la frontière du maillage $\bar{\omega}$, tandis que

$$y_{\bar{x}_1 x_1} = \frac{1}{h_1^2}(y(i+1, j) - 2y(i, j) + y(i-1, j)),$$

$$y_{\bar{x}_2 x_2} = \frac{1}{h_2^2}(y(i, j+1) - 2y(i, j) + y(i, j-1)),$$

$$y(i, j) = y(x_{ij}).$$

Au § 2, ch. V, on a montré que le problème de différences (4) peut être réduit à l'équation opératorielle (1) pour laquelle l'opérateur A se détermine de la façon suivante: $Ay = -\Lambda \overset{\circ}{y}$, où $y \in H$, $\overset{\circ}{y} \in \overset{\circ}{H}$ et $\overset{\circ}{y}(x) = y(x)$ pour $x \in \omega$. $\overset{\circ}{H}$ est ici un ensemble des fonctions de mailles associées à $\bar{\omega}$ et s'annulant sur γ, tandis que H est l'espace de fonctions de mailles données sur ω et dont le produit scalaire $(u, v) = \sum_{x \in \omega} u(x) v(x) h_1 h_2$. Le second membre f de (1) ne diffère du second membre φ de l'équation aux différences (4) que dans les nœuds frontières:

$$f(x) = \varphi(x) + \varphi_1(x)/h_1^2 + \varphi_2(x)/h_2^2,$$

$$\varphi_1(x) = \begin{cases} g(0, x_2), & x_1 = h_1, \\ 0, & 2h_1 \leqslant x_1 \leqslant l_1 - 2h_1, \\ g(l_1, x_2), & x_1 = l_1 - h_1, \end{cases}$$

$$\varphi_2(x) = \begin{cases} g(x_1, 0), & x_2 = h_2, \\ 0, & 2h_2 \leqslant x_2 \leqslant l_2 - 2h_2, \\ g(x_1, l_2), & x_2 = l_2 - h_2. \end{cases}$$

Bref, on a réduit le problème aux limites discret (4) à l'équation opératorielle (1) dans un espace hilbertien de dimension finie H.

Pour la résolution approchée de l'équation (1), prenons la *méthode explicite de Tchébychev* ($B = E$):

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \ldots, \quad y_0 \in H, \tag{5}$$

$$\tau_k = \frac{\tau_0}{1 + \rho_0 \mu_k}, \quad \mu_k \in \mathfrak{M}_n^* = \left\{ -\cos\frac{(2i-1)\pi}{2n}, \quad i = 1, 2, \ldots, n \right\}, \tag{6}$$

$$k = 1, 2, \ldots, n,$$

$$n \geqslant n_0(\varepsilon), \quad n_0(\varepsilon) = \ln(0{,}5\varepsilon)/\ln \rho_1. \tag{7}$$

Au § 2, ch. V, on a montré que l'opérateur A qu'on vient de définir est autoadjoint dans H et ses bornes γ_1 et γ_2 coïncident avec les valeurs propres minimale et maximale de l'opérateur de différences Λ. c'est-à-dire

$$\gamma_1 E \leqslant A \leqslant \gamma_2 E, \quad \gamma_1 > 0, \tag{8}$$

où

$$\gamma_1 = \sum_{\alpha=1}^{2} \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2l_\alpha}, \quad \gamma_2 = \sum_{\alpha=1}^{2} \frac{4}{h_\alpha^2} \cos^2 \frac{\pi h_\alpha}{2l_\alpha}. \tag{9}$$

Les opérateurs A et $B = E$ sont autoadjoints et définis positifs dans H. Il s'ensuit donc des exemples examinés au point 3 du § 2 que γ_1, γ_2 de (8) sont des constantes de la méthode de Tchébychev (5)-(7), si en guise de D est choisi l'un des opérateurs E, A ou A^2. Alors dans les formules (6). (7)

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2},$$

où γ_1 et γ_2 sont définis dans (9).

Vu que $\gamma_1 = O(1)$, tandis que $\gamma_2 = O(1/h_1^2 + 1/h_2^2)$, on a $\xi = O(|h|^2)$, où $|h|^2 = h_1^2 + h_2^2$. Donc pour l'exemple étudié l'estimation asymptotique du nombre d'itérations $n_0(\varepsilon)$ prend la forme

$$n_0(\varepsilon) = O\left(\frac{1}{|h|} \ln \frac{2}{\varepsilon}\right).$$

Dans le cas particulier de $\bar{G}$ carré de côté l ($l_1 = l_2 = l$) et de maillage $\bar{\omega}$ carré ($h_1 = h_2 = h = l/N$), on a

$$\gamma_1 = \frac{8}{h^2} \sin^2 \frac{\pi h}{2l}, \quad \gamma_2 = \frac{8}{h^2} \cos^2 \frac{\pi h}{2l}, \quad \xi = \operatorname{tg}^2 \frac{\pi h}{2l},$$

$$\tau_0 = \frac{h^2}{4}, \quad \rho_0 = \cos \frac{\pi h}{l}, \quad \rho_1 = \frac{1 - \sin \frac{\pi h}{l}}{\cos \frac{\pi h}{l}},$$

$$n_0(\varepsilon) \approx \frac{l}{\pi h} \ln \frac{2}{\varepsilon} \approx 0{,}32N \ln \frac{2}{\varepsilon}. \tag{10}$$

Donc, le nombre d'itérations n est proportionnel au nombre de nœuds N dans une direction. Notons que le nombre d'inconnues du problème (4) est égal à $M = (N-1)^2$, c'est-à-dire que le nombre d'itérations est proportionnel à la racine carrée du nombre d'inconnues.

Le schéma opératoriel itératif (5) pour $B = E$ peut être écrit, en utilisant la définition de l'opérateur A et du second membre f, sous forme de schéma aux différences suivant:

$$y_{k+1} = y_k + \tau_{k+1}(\Lambda y_k + \varphi), \quad x \in \omega, \quad y_k|_\gamma = g, \quad k = 0, 1, \ldots$$

En y portant (4), on obtient les formules de calcul

$$y_{k+1}(i,j)=\left(1-\frac{\tau_{k+1}}{\tau_0}\right)y_k(i,j)+$$
$$+\tau_{k+1}\left[\frac{y_k(i+1,j)+y_k(i-1,j)}{h_1^2}+\frac{y_k(i,j+1)+y_k(i,j-1)}{h_2^2}+\varphi(i,j)\right],$$
$$1\leqslant i\leqslant N_1-1,\quad 1\leqslant j\leqslant N_2-1.$$

L'approximation initiale y_0 est une fonction de maille quelconque sur ω prenant à la frontière γ les valeurs données $y_0(x)=g(x)$ pour $x\in\gamma$.

Apprécions le nombre d'opérations arithmétiques $Q(\varepsilon)$ nécessaire à l'obtention de la solution approchée du problème de différences (4) à la précision ε en utilisant la méthode de Tchébychev (5)-(7).

En posant donnés les paramètres d'itération τ_k, on obtient que pour le calcul de y_{k+1} dans un nœud du maillage ω il faut neuf opérations arithmétiques. Le nombre de nœuds intérieurs dans le maillage $\bar{\omega}$ étant $M=(N_1-1)(N_2-1)$, la réalisation d'une itération exige $Q_0\approx 9N_1N_2$ opérations arithmétiques. Donc $Q(\varepsilon)=nQ_0\approx$ $\approx 9nN_1N_2$, où n est le nombre d'itérations.

Pour le cas particulier examiné plus haut le nombre d'itérations n est déterminé dans (10) et, par suite, cet exemple donne

$$Q(\varepsilon)\approx 2.9N^3\ln(2/\varepsilon).$$

Pour résoudre l'équation (1), voyons maintenant la *méthode itérative simple*. Le schéma itératif de la méthode itérative simple a la forme (5), tandis que les paramètres d'itération τ_k et le nombre d'itérations n s'obtiennent à l'aide des formules du théorème 2:

$$\tau_k\equiv\tau_0=\frac{2}{\gamma_1+\gamma_2},\quad n\geqslant n_0(\varepsilon)=\frac{\ln\varepsilon}{\ln\rho_0},\quad \rho_0=\frac{1-\xi}{1+\xi},\quad \xi=\frac{\gamma_1}{\gamma_2},\tag{11}$$

où γ_1 et γ_2 sont définis dans (9). A partir de (9) et (11) on obtient l'estimation asymptotique en h du nombre d'itérations de la méthode itérative simple $n_0(\varepsilon)=O\left(\frac{1}{|h^2|}\ln\frac{1}{\varepsilon}\right)$. Pour le cas particulier étudié plus haut, on obtient

$$n_0(\varepsilon)\approx\frac{2l^2}{\pi^2h^2}\ln\frac{1}{\varepsilon}\approx 0{,}2N^2\ln\frac{1}{\varepsilon},\tag{12}$$

autrement dit, le nombre d'itérations de la méthode itérative simple est proportionnel au carré des nœuds N suivant une direction (ou au nombre d'inconnues de l'équation).

En comparant les estimations du nombre d'itérations des méthodes de Tchébychev (10) et itérative simple (12) on constate que la méthode itérative simple exige un nombre beaucoup plus important d'itérations que la méthode de Tchébychev. Pour la confrontation de ces méthodes sur des maillages réels, fournissons les valeurs vraies du nombre d'itérations pour le cas particulier susmentionné

en fonction du nombre de nœuds N suivant une direction pour $\varepsilon = 10^{-4}$ (en premier lieu on indique le nombre d'itérations de la méthode de Tchébychev):

$$\begin{array}{lll} N = 32 & n = 101 & n = 1909 \\ N = 64 & n = 202 & n = 7642 \\ N = 128 & n = 404 & n = 30577. \end{array}$$

Fournissons les formules de calcul de la méthode itérative simple pour le cas particulier étudié:

$$y_{k+1}(i, j) = \frac{1}{4}\Big[y_k(i+1, j) + y_k(i-1, j) + y_k(i, j+1) + y_k(i, j-1)\Big] + \frac{h^2}{4}\varphi(i, j).$$

La très forte dépendance du nombre d'itérations de la méthode itérative simple du nombre de nœuds N du maillage est la raison de ce qu'actuellement cette méthode n'est presque pas utilisée pour la résolution des équations de mailles elliptiques.

2. Problème de différences de Dirichlet pour l'équation de Poisson dans un domaine quelconque. Supposons qu'il s'agit dans un domaine limité quelconque G à frontière Γ de rechercher la solution de l'équation de Poisson

$$Lu = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -f(x), \quad x = (x_1, x_2) \in G, \tag{13}$$

qui prend à la frontière Γ les valeurs données

$$u(x) = g(x), \quad x \in \Gamma. \tag{14}$$

A titre de simplification, examinons le cas quand l'intersection du domaine G avec la droite passant par le point $x \in G$ et parallèle à l'axe des coordonnées n'intercepte qu'un seul intervalle.

Recouvrons le plan d'un réseau formé par l'intersection de droites parallèles aux axes de coordonnées et menées à la même distance h l'une de l'autre.

Les points x_{ij} du réseau appartenant à G appelons *nœuds* du maillage $\omega = \{x_{ij} \in G\}$. Désignons par $\Delta_\alpha(x_{ij})$ l'intervalle formé par l'intersection de G avec la droite passant par le point $x_{ij} \in \omega$ parallèlement à l'axe des coordonnées Ox_α, $\alpha = 1, 2$. Les extrémités de cet intervalle seront appelées *nœuds frontières* dans la direction x_α. L'ensemble de tous les nœuds frontières dans la direction x_α sera désigné par γ_α, tandis que par $\gamma = \gamma_1 \cup \gamma_2$ sera désignée la frontière du domaine maillé. Les ensembles des nœuds intérieurs et frontières constituent le maillage $\bar{\omega} = \omega \cup \gamma$ dans le domaine $\bar{G}$.

Prenons l'un des intervalles Δ_α. L'ensemble de nœuds $x_{ij} \in \omega$ se plaçant sur cet intervalle sera désigné par $\omega_\alpha(x_\beta)$, $\beta = 3 - \alpha$, $\alpha = 1, 2$. Au moyen de $\omega_\alpha^+(x_\beta)$ désignons l'ensemble composé de

nœuds $\omega_\alpha (x_\beta)$ et de l'extrémité droite de l'intervalle Δ_α. Définissons $\bar{\omega}_\alpha (x_\beta)$ comme l'ensemble composé de nœuds $\omega_\alpha (x_\beta)$ et d'extrémités de l'intervalle Δ_α. Désignons par $x_{ij}^{(+1_\alpha)}$ et $x_{ij}^{-1_\alpha}$ les nœuds voisins du point $x_{ij} \in \omega_\alpha (x_\beta)$ respectivement à droite et à gauche et appartenant à $\bar{\omega}_\alpha (x_\beta)$.

On appellera *pas* $h_\alpha^\pm (x_{ij})$ du maillage ω au point $x_{ij} \in \omega$ la distance entre les nœuds x_{ij} et $x_{ij}^{(\pm 1_\alpha)} \in \bar{\omega}$. Notons que si tous les quatre nœuds $x_{ij}^{(\pm 1_\alpha)}$ voisins de x_{ij} appartiennent à ω les pas $h_\alpha^\pm$ sont alors égaux au pas principal du réseau h. Dans les nœuds voisins de la frontière $h_\alpha^\pm \leqslant h$. Entre les pas h_α^+ et h_α^- on a la relation

$$h_\alpha^+ (x_{ij}) = h_\alpha^- (x_{ij}^{(+1_\alpha)}).$$

Au problème (13), (14) faisons correspondre sur le maillage $\bar{\omega}$ le problème aux limites discret

$$\Lambda y = \sum_{\alpha=1}^{2} y_{\bar{x}_\alpha \hat{x}_\alpha} = -\varphi(x), \quad x \in \omega, \quad y(x) = g(x), \quad x \in \gamma, \qquad (15)$$

où

$$y_{\bar{x}_\alpha} = \frac{1}{h_\alpha^-} (y - y^{-1_\alpha}), \quad y_{x_\alpha} = \frac{1}{h_\alpha^+} (y^{+1_\alpha} - y),$$

$$y_{\hat{x}_\alpha} = \frac{1}{h} (y^{+1_\alpha} - y), \quad y_{\bar{x}_\alpha \hat{x}_\alpha} = \frac{1}{h} \left(\frac{y^{+1_\alpha} - y}{h_\alpha^+} - \frac{y - y^{-1_\alpha}}{h_\alpha^-} \right),$$

$$y^{\pm 1_\alpha} = y(x^{(\pm 1_\alpha)}), \qquad \alpha = 1, 2.$$

Le problème de différences (15) se réduit à l'équation opératorielle (1), l'opérateur A se définissant de la même façon qu'au point 1. Le produit scalaire dans H est donné ainsi:

$$(u, v) = \sum_{x \in \omega} u(x) v(x) h^2.$$

Introduisons quelques notations qu'on utilisera dans la suite. Définissons les produits scalaires pour les fonctions de mailles données sur $\bar{\omega}$ au moyen des formules:

$$(u, v)_{\omega_\alpha} = \sum_{x_\alpha \in \omega_\alpha (x_\beta)} u(x) v(x) h,$$

$$(u, v)_{\omega_\alpha^+} = \sum_{x_\alpha \in \omega_\alpha^+ (x_\beta)} u(x) v(x) h_\alpha^- (x),$$

$$(u, v)_\alpha = ((u, v)_{\omega_\alpha^+}, 1)_{\omega_\beta}, \quad \beta = 3 - \alpha, \quad \alpha = 1, 2.$$

En recourant à ces notations, on peut écrire le produit scalaire dans H sous la forme

$$(u, v) = ((u, v)_{\omega_1}, 1)_{\omega_2} = ((u, v)_{\omega_2}, 1)_{\omega_1}. \qquad (16)$$

De la définition de l'opérateur A on déduit que

$$(Au, v) = -(\Lambda \overset{\circ}{u}, \overset{\circ}{v}) = -((\overset{\circ}{u}_{\bar{x}_1\hat{x}_1}, \overset{\circ}{v})_{\omega_1}, 1)_{\omega_2} - ((\overset{\circ}{u}_{\bar{x}_2\hat{x}_2}, \overset{\circ}{v})_{\omega_2}, 1)_{\omega_1}.$$

Or, comme en vertu des formules de différences de Green on a l'égalité (voir point 3, § 2, ch. V)

$$-(\overset{\circ}{u}_{\bar{x}_\alpha\hat{x}_\alpha}, \overset{\circ}{v})_{\omega_\alpha} = (\overset{\circ}{u}_{\bar{x}_\alpha}, \overset{\circ}{v}_{\bar{x}_\alpha})_{\omega_\alpha^+} = -(\overset{\circ}{u}, \overset{\circ}{v}_{\bar{x}_\alpha\hat{x}_\alpha})_{\omega_\alpha},$$

on en déduit les égalités

$$(Au, v) = (u, Av),$$

$$(Au, u) = \sum_{\alpha=1}^{2} (\overset{\circ}{u}^2_{\bar{x}_\alpha}, 1)_\alpha, \quad u \in H, \quad \overset{\circ}{u} \in \overset{\circ}{H}. \tag{17}$$

La première de ces égalités démontre que l'opérateur A est autoadjoint dans H.

Pour résoudre de façon approchée l'équation (1), adressons-nous à la *méthode implicite de Tchébychev*

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \ldots, \quad y_0 \in H,$$

où en guise d'opérateur B est pris l'opérateur diagonal facilement inversible

$$By = (b_1 + b_2)\, y, \quad b_\alpha(x) = \frac{1}{h}\left(\frac{1}{h_\alpha^+(x)} + \frac{1}{h_\alpha^-(x)}\right), \quad \alpha = 1, 2. \tag{18}$$

Expliquons le choix de l'opérateur B. Si l'équation (1) est traitée comme un système d'équations algébriques linéaires de matrice $\mathcal{A}$ correspondant à l'opérateur A, la matrice $\mathcal{B}$ correspondant à l'opérateur B est alors la partie diagonale de la matrice $\mathcal{A}$.

Comme les opérateurs A et B sont autoadjoints et définis positifs dans H, γ_1 et γ_2 compris dans les conditions (6), (7) constituent des constantes de l'équivalence énergétique des opérateurs A et B:

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \quad \gamma_1 > 0,$$

si en guise de D est choisi l'un des opérateurs A, B ou $AB^{-1}A$.

Cherchons les estimations pour γ_1 et γ_2. Montrons d'abord qu'on a l'égalité

$$\gamma_1 + \gamma_2 = 2. \tag{19}$$

En effet, soit $u\ (x)$ la fonction de maille arbitraire de H. Examinons la fonction $v\ (x)$ qu'on définira de la façon suivante:

$$v\ (x_{ij}) = (-1)^{i+j}\, u\ (x_{ij}), \quad x_{ij} \in \omega.$$

Calculons la valeur de l'opérateur de différences $\Lambda\overset{\circ}{v}$ au point x_{ij}. Il vient

$$\Lambda\overset{\circ}{v}(i, j)=\sum_{\alpha=1}^{2}\frac{1}{h}\left(\frac{\overset{\circ}{v}^{+1_\alpha}-\overset{\circ}{v}}{h_\alpha^+}-\frac{\overset{\circ}{v}-\overset{\circ}{v}^{-1_\alpha}}{h_\alpha^-}\right)_{x=x_{ij}}=$$

$$=-(-1)^{i+j}\sum_{\alpha=1}^{2}\frac{1}{h}\left(\frac{\overset{\circ}{u}^{+1_\alpha}-\overset{\circ}{u}}{h_\alpha^+}-\frac{\overset{\circ}{u}-\overset{\circ}{u}^{-1_\alpha}}{h_\alpha^-}\right)_{x=x_{ij}}-$$

$$-2(-1)^{i+j}(b_1(x_{ij})+b_2(x_{ij}))\overset{\circ}{u}(x_{ij}).$$

Donc

$$Av(i, j) = -\Lambda\overset{\circ}{v}(i, j) = (-1)^{i+j}(2B - A)u(i, j).$$

Ensuite, vu que

$$\gamma_1=\min_{u\neq 0}\frac{(Au, u)}{(Bu, u)},\quad (Av, v)=2(Bu, u)-(Au, u),\quad (Bv, v)=(Bu, u),$$

on a

$$\gamma_2=\max_{v\neq 0}\frac{(Av, v)}{(Bv, v)}=2-\min_{u\neq 0}\frac{(Au, u)}{(Bu, u)}=2-\gamma_1.$$

La proposition est démontrée.

En utilisant la relation (19) on obtient que dans les formules (6)

$$\tau_0 = 2/(\gamma_1 + \gamma_2) = 1,\quad \rho_0 = (\gamma_2 - \gamma_1)/(\gamma_2 + \gamma_1) = 1 - \gamma_1.$$

Pour le calcul des paramètres d'itération τ_k il suffit donc de trouver l'estimation pour γ_1. A partir du lemme 13, § 2, ch. V, il s'ensuit que pour toute fonction de maille $\overset{\circ}{y}\in\overset{\circ}{H}$ on a l'inégalité

$$(b_\alpha\overset{\circ}{y}, \overset{\circ}{y})_{\omega_\alpha}\leqslant\varkappa_\alpha(\overset{\circ}{y}{}^2_{\bar{x}_\alpha}, 1)_{\omega_\alpha^+},\quad \alpha=1, 2, \tag{20}$$

où $\varkappa_\alpha=\varkappa_\alpha(x_\beta)=\max\limits_{x_\alpha\in\omega_\alpha(x_\beta)} v^\alpha(x)$, quant à $v^\alpha(x)$, c'est la solution du problème aux limites triponctuel suivant:

$$\begin{aligned}&v^\alpha_{\bar{x}_\alpha\hat{x}_\alpha}=-b_\alpha(x),\quad &&x_\alpha\in\omega_\alpha(x_\beta),\\ &v^\alpha(x)=0, &&x_\alpha\in\gamma_\alpha.\end{aligned}\tag{21}$$

En divisant l'inégalité (20) par $\varkappa_\alpha$ et en sommant en ω_β, il vient

$$\left(\frac{b_\alpha}{\varkappa_\alpha}\overset{\circ}{y}, \overset{\circ}{y}\right)\leqslant((\overset{\circ}{y}{}^2_{\bar{x}_\alpha}, 1)_{\omega_\alpha^+}, 1)_{\omega_\beta}=(\overset{\circ}{y}{}^2_{\bar{x}_\alpha}, 1)_\alpha,\quad \alpha=1, 2.$$

En additionnant ces inégalités, on obtient

$$\left(\sum_{\alpha=1}^{2}\frac{b_\alpha}{\varkappa_\alpha}\overset{\circ}{y}, \overset{\circ}{y}\right)\leqslant\sum_{\alpha=1}^{2}(\overset{\circ}{y}{}^2_{\bar{x}_\alpha}, 1)_\alpha. \tag{22}$$

h_α^- h h_α^+

$l_\alpha(x_\beta)$ $L_\alpha(x_\beta)$

Fig. 3.

De (17), (18) et (22) il s'ensuit que pour γ_1 on peut choisir la quantité

$$\gamma_1 = \min_{x \in \omega} \frac{1}{b_1(x) + b_2(x)} \sum_{\alpha=1}^{2} \frac{b_\alpha(x)}{\varkappa_\alpha(x_\beta)}. \qquad (23)$$

Il ne reste qu'à calculer $\varkappa_\alpha$. Cherchons pour cela la solution du problème (21).

Soient $l_\alpha(x_\beta)$ et $L_\alpha(x_\beta)$ les extrémités de l'intervalle Δ_α sur lequel se disposent les nœuds du maillage $\omega_\alpha(x_\beta)$. En vertu de la construction du réseau sur le plan les pas $h_\alpha^\pm$ ne diffèrent de h que pour les nœuds voisins de la frontière (voir fig. 3).

Aussi sur le maillage $\omega_\alpha(x_\beta)$ la différence divisée $v_{\bar{x}_\alpha \hat{x}_\alpha}$ et le second membre de l'équation (21) peuvent-ils s'écrire de la sorte

$$v_{\bar{x}_\alpha \hat{x}_\alpha} = \frac{1}{h}\left(\frac{v^{+1_\alpha} - v}{h} - \frac{v - v^{-1_\alpha}}{h_\alpha^-}\right), \quad b_\alpha = \frac{1}{h}\left(\frac{1}{h} + \frac{1}{h_\alpha^-}\right), \quad x_\alpha = l_\alpha + h_\alpha^-,$$

$$v_{\bar{x}_\alpha \hat{x}_\alpha} = \frac{1}{h^2}(v^{+1_\alpha} - 2v + v^{-1_\alpha}), \qquad b_\alpha = \frac{2}{h^2}, \quad l_\alpha + h_\alpha^- < x_\alpha < L_\alpha - h_\alpha^+,$$

$$v_{\bar{x}_\alpha \hat{x}_\alpha} = \frac{1}{h}\left(\frac{v^{+1_\alpha} - v}{h_\alpha^+} - \frac{v - v^{-1_\alpha}}{h}\right), \quad b_\alpha = \frac{1}{h}\left(\frac{1}{h} + \frac{1}{h_\alpha^+}\right), \quad x_\alpha = L_\alpha - h_\alpha^+.$$

La vérification directe montre que la fonction de maille

$$v^\alpha(x) = \frac{1}{h^2}\left[(x_\alpha - l_\alpha)\left(L_\alpha - x_\alpha + \frac{(h_\alpha^-)^2 - (h_\alpha^+)^2}{L_\alpha - l_\alpha}\right) + h^2 - (h_\alpha^-)^2\right]$$

pour $x_\alpha \in \omega_\alpha(x_\beta)$ est la solution du problème (21). Vu que

$$v^\alpha(x) \leqslant \frac{1}{h^2}(x_\alpha - l_\alpha)(L_\alpha - x_\alpha) + 1,$$

on a

$$\varkappa_\alpha = \max_{x_\alpha \in \omega_\alpha(x_\beta)} v^\alpha(x) \leqslant \frac{1}{h^2}\left(\frac{L_\alpha - l_\alpha}{2}\right)^2 + 1. \qquad (24)$$

En portant (18) et (24) dans (23) on obtient l'estimation pour γ_1.

L'estimation grossière de γ_1 peut être obtenue de la façon suivante. Supposons que le domaine étudié $\bar{G}$ soit inscrit dans un carré de côté l. Dans ce cas $L_\alpha - l_\alpha \leqslant l$ pour tout α et, partant, $\varkappa_\alpha \leqslant$

$\leqslant l^2/(4h^2) + 1$, $\alpha = 1, 2$. En portant cette estimation dans (23), on obtient $\gamma_1 \geqslant 4h^2/(l^2 + 4h^2)$, c'est-à-dire que $\gamma_1 \approx 4h^2/l^2$. Comme $\gamma_2 = 2 - \gamma_1$, $\xi = \gamma_1/\gamma_2 \approx 2h^2/l^2$. Par conséquent, de l'estimation (7) du nombre d'itérations on tire

$$n_0(\varepsilon) \approx \frac{l^2}{2\sqrt{2}h} \ln \frac{2}{\varepsilon} \approx 0{,}35N \ln \frac{2}{\varepsilon}, \tag{25}$$

où N est le nombre maximal de nœuds en chaque direction.

Donc, pour la méthode de Tchébychev implicite étudiée ici le nombre d'itérations ne dépend que du pas principal du maillage h et est indépendant des pas irréguliers $h_\alpha^\pm$ dans les nœuds voisins de la frontière. En comparant l'estimation (25) à celle obtenue auparavant dans (10), on constate que le nombre d'itérations pour le cas du domaine arbitraire $\bar{G}$ inscrit dans un carré de côté l est le même que pour le cas où le domaine $\bar{G}$ est précisément le carré mentionné.

Donnons les formules de calcul pour la méthode itérative de Tchébychev quand cette dernière est utilisée pour la résolution du problème de différences de Dirichlet pour l'équation de Poisson dans un domaine $\bar{G}$ arbitraire:

$$y_{k+1}(x_{ij}) = (1 - \tau_{k+1})\, y_k(x_{ij}) + \\ + \frac{\tau_{k+1}}{b_1(x_{ij}) + b_2(x_{ij})} \left[\frac{1}{h}\left(\frac{y^{+1_1}}{h_1^+} + \frac{y^{-1_1}}{h_1^-} + \frac{y^{+1_2}}{h_2^+} + \frac{y^{-1_2}}{h_2^-}\right) + \varphi\right]_{x=x_{ij}},$$

$$x_{ij} \in \omega, \quad y_k(x) = g(x), \quad x \in \gamma.$$

Notons que dans le ch. X il sera construit pour le problème étudié une autre méthode implicite de Tchébychev (méthode itérative triangulaire alternée) pour laquelle le nombre d'opérations arithmétiques de la mise en œuvre d'une itération est quelque peu supérieur à celui que nécessite la méthode qu'on vient de décrire, quant au nombre d'itérations, il est sensiblement inférieur, ce qui rend la méthode très efficace.

3. Problème de différences de Dirichlet pour l'équation elliptique à coefficients variables.

3.1. *Méthode explicite de Tchébychev.* Etudions le problème de Dirichlet pour l'équation elliptique de second ordre à coefficients variables dans le rectangle $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha,\ \alpha = 1, 2\}$:

$$Lu = \sum_{\alpha=1}^{2} \frac{\partial}{\partial x_\alpha}\left(k_\alpha(x) \frac{\partial u}{\partial x_\alpha}\right) - q(x)\, u = -f(x), \quad x \in G, \\ u(x) = g(x), \quad x \in \Gamma. \tag{26}$$

Sur le maillage rectangulaire

$$\bar{\omega}=\{x_{ij}=(ih_1,\ jh_2)\in\bar{G},\quad 0\leqslant i\leqslant N_1,\quad 0\leqslant j\leqslant N_2,\quad h_\alpha=l_\alpha/N_\alpha,\quad \alpha=1,\ 2\}$$

au problème différentiel (26) correspond le problème de différences

$$\Lambda y=\sum_{\alpha=1}^{2}(a_\alpha(x)\,y_{\bar{x}_\alpha})_{x_\alpha}-d(x)\,y=-\varphi(x),\quad x\in\omega, \qquad y(x)=g(x),\quad x\in\gamma. \tag{27}$$

Si les coefficients $k_\alpha(x)$, $q(x)$ et $f(x)$ constituent des fonctions suffisamment lisses, les coefficients $a_\alpha(x)$, $d(x)$ et $\varphi(x)$ du schéma aux différences (27) peuvent alors être définis, par exemple, de la façon suivante:

$$a_1(x_{ij})=k_1((i-0{,}5)\,h_1, jh_2),\quad a_2(x_{ij})=k_2(ih_1,\ (j-0{,}5)\,h_2),$$
$$d(x_{ij})=q(x_{ij}),\qquad \varphi(x_{ij})=f(x_{ij}).$$

Admettons que les coefficients du schéma aux différences (27) satisfont aux conditions

$$0<c_1\leqslant a_\alpha(x)\leqslant c_2,\quad x\in\bar{\omega}, \qquad 0\leqslant d_1\leqslant d(x)\leqslant d_2,\quad x\in\omega,\quad \alpha=1,\ 2. \tag{28}$$

Ces conditions garantissent l'existence et l'unicité de la solution du problème (27).

Le schéma aux différences (27) se réduit à l'équation opératorielle (1) de la façon banale: $Ay=-\Lambda\overset{\circ}{y}$, où $y\in H$, $\overset{\circ}{y}\in\overset{\circ}{H}$, tandis que H est l'espace de fonctions de mailles associées à ω au produit scalaire

$$(u,\ v)=\sum_{x\in\omega}u(x)\,v(x)\,h_1h_2,$$

le second membre f de l'équation (1) ne diffère du second membre φ du schéma (27) que dans les nœuds voisins de la frontière.

Pour résoudre de façon approchée l'équation (1), recourrons à la méthode explicite de Tchébychev (5)-(7) ($B=E$). On a montré au § 2, ch. V, que l'opérateur A défini ici est autoadjoint dans H. Aussi l'information à priori de la méthode de Tchébychev acquiert-elle la forme de constantes γ_1 et γ_2 des inégalités $\gamma_1E\leqslant A\leqslant\gamma_2E$, $\gamma_1>0$, si en qualité de D on a choisi l'un des opérateurs E, A ou A^2. Cherchons ces constantes. A cette fin introduisons l'opérateur $\overset{\circ}{A}$, correspondant à l'opérateur de différences $\overset{\circ}{\Lambda}$, où $\overset{\circ}{\Lambda}y=y_{\bar{x}_1x_1}+y_{\bar{x}_2x_2}$, et définissons les produits scalaires suivants pour les fonctions de mailles données sur $\bar{\omega}$:

$$(u,\ v)_{\omega_\alpha}=\sum_{x_\alpha\in\omega_\alpha}u(x)\,v(x)\,h_\alpha,\quad (u,\ v)_{\omega_\alpha^+}=\sum_{x_\alpha\in\omega_\alpha^+}u(x)\,v(x)\,h_\alpha,$$
$$(u,\ v)_\alpha=((u,\ v)_{\omega_\alpha^+},\ 1)_{\omega_\beta},\quad \beta=3-\alpha,\quad \alpha=1,\ 2.$$

On a ici

$$\omega_\alpha=\{x_{\alpha,i}=ih_\alpha,\quad 1\leqslant i\leqslant N_\alpha-1\},\quad \omega_\alpha^+=\{x_{\alpha,i}=ih_\alpha,\quad 1\leqslant i\leqslant N_\alpha\}.$$

Les produits scalaires qu'on a introduits ici sont des analogues des produits scalaires définis au point 2.

A partir de la définition des opérateurs A et $\mathring{A}$ et des formules aux différences de Green (voir point 2, § 2, ch. V) on obtient

$$(Au,u)=-(\Lambda\mathring{u},\mathring{u})=-((a_1\mathring{u}_{\bar{x}_1x_1},\mathring{u})_{\omega_1},1)_{\omega_2}-((a_2\mathring{u}_{\bar{x}_2x_2},\mathring{u})_{\omega_2},1)_{\omega_1}+$$

$$+(du,u)=\sum_{\alpha=1}^{2}(a_\alpha\mathring{u}^2_{\bar{x}_\alpha},1)_\alpha+(du,u),\quad (29)$$

$$(\mathring{A}u,u)=-(\mathring{\Lambda}\mathring{u},\mathring{u})=\sum_{\alpha=1}^{2}(\mathring{u}^2_{\bar{x}_\alpha},1)_\alpha,\quad u\in H,\quad \mathring{u}\in \mathring{H}.$$

Compte tenu des inégalités (28), on obtient de ce qui précède les inégalités opératorielles de la forme

$$c_1\mathring{A}+d_1E\leqslant A\leqslant c_2\mathring{A}+d_2E.\quad (30)$$

Au point 1, § 5, il a été montré que l'opérateur $\mathring{A}$ possède des bornes

$$\mathring{\gamma}_1=\sum_{\alpha=1}^{2}\frac{4}{h_\alpha^2}\sin^2\frac{\pi h_\alpha}{2l_\alpha},\quad \mathring{\gamma}_2=\sum_{\alpha=1}^{2}\frac{4}{h_\alpha^2}\cos^2\frac{\pi h_\alpha}{2l_\alpha},$$

c'est-à-dire qu'on a les inégalités

$$\mathring{\gamma}_1E\leqslant\mathring{A}\leqslant\mathring{\gamma}_2E.\quad (31)$$

A partir de (30) et (31) on obtient que l'opérateur A possède des bornes $\gamma_1=c_1\mathring{\gamma}_1+d_1$, $\gamma_2=c_2\mathring{\gamma}_2+d_2$.

Bref, les constantes γ_1 et γ_2 sont trouvées. En les utilisant, calculons suivant les formules (6) les paramètres d'itération τ_k, tandis qu'à l'aide des formules (7) cherchons l'estimation pour le nombre d'itérations n.

Vu que $\mathring{\xi}=\mathring{\gamma}_1/\mathring{\gamma}_2=O(|h|^2)$, $\xi=\gamma_1/\gamma_2=O(|h|^2)$ et on a pour le nombre d'itérations de la méthode étudiée l'estimation asymptotique suivante:

$$n_0(\varepsilon)=O\left(\frac{1}{|h|}\ln\frac{2}{\varepsilon}\right),$$

tandis que la constante dans l'estimation est fonction des caractéristiques extrémales des coefficients $a_\alpha(x)$ et $d(x)$, autrement dit, de c_α et d_α, $\alpha=1, 2$.

Dans le cas particulier, quand $\overline{G}$ est un carré de côté l ($l_1 = l_2 = l$), le maillage $\overline{\omega}$ est carré ($h_1 = h_2 = h = l/N$) et $d \equiv 0$, on obtient

$$\gamma_1 = \frac{8c_1}{h^2} \sin^2 \frac{\pi h}{2l}, \quad \gamma_2 = \frac{8c_2}{h^2} \cos^2 \frac{\pi h}{2l}, \quad \xi = \frac{c_1}{c_2} \operatorname{tg}^2 \frac{\pi h}{2l}$$

et, partant,

$$n_0(\varepsilon) \approx \sqrt{\frac{c_2}{c_1}} \frac{l}{\pi h} \ln \frac{2}{\varepsilon} \approx 0{,}32 \sqrt{\frac{c_2}{c_1}} N \ln \frac{2}{\varepsilon}.$$

En comparant l'estimation obtenue pour le nombre d'itérations de la méthode de résolution explicite de Tchébychev de l'équation aux différences (27) à coefficients variables avec l'estimation (10), on trouve que pour l'exemple considéré le nombre d'itérations est $\sqrt{c_2/c_1}$ fois supérieur au nombre d'itérations quand les coefficients sont constants.

Donnons les formules de calcul pour la méthode explicite de Tchébychev (5)-(7) utilisée à la résolution de l'équation aux différences (27). Ces formules prennent la forme:

$$y_{k+1}(i, j) = \alpha_{k+1}(i, j)\, y_k(i, j) + \\ + \tau_{k+1}\Big\{\frac{1}{h_1^2}[a_1(i+1, j)\, y_k(i+1, j) + a_1(i, j)\, y_k(i-1, j)] + \\ + \frac{1}{h_2^2}[a_2(i, j+1)\, y_k(i, j+1) + a_2(i, j)\, y_k(i, j-1)] + \varphi(i, j)\Big\},$$
$$1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant j \leqslant N_2 - 1,$$

où on a posé

$$\alpha_{k+1}(i, j) = 1 - \tau_{k+1}\Big[\frac{a_1(i+1, j) + a_1(i, j)}{h_1^2} + \\ + \frac{a_2(i, j+1) + a_2(i, j)}{h_2^2} + d(i, j)\Big],$$

tandis que l'estimation initiale y_0 est une fonction de maille arbitraire sur ω qui prend sur γ les valeurs fixées: $y(x) = g(x)$ pour $x \in \gamma$.

3.2. *Méthode implicite de Tchébychev.* Pour résoudre de façon approchée l'équation (1) construite au point précédent et correspondant au schéma discret (27), voyons maintenant la méthode implicite la plus simple de Tchébychev (5)-(7). En guise d'opérateur B prenons de nouveau, comme au point 2, la partie diagonale de l'opérateur A

$$By = by,$$
$$b(i, j) = \frac{1}{h_1^2}[a_1(i+1, j) + a_1(i, j)] + \\ + \frac{1}{h_2^2}[a_2(i, j+1) + a_2(i, j)] + d(i, j). \quad (32)$$

Etant donné que les opérateurs A et B sont autoadjoints dans H et définis positifs, l'information à priori de la méthode implicite de Tchébychev (5)-(7) acquiert l'aspect de constantes de l'équivalence énergétique des opérateurs $\gamma_1 B \leqslant A \leqslant \gamma_2 B$, $\gamma_1 > 0$, si en guise de D on a choisi l'un des opérateurs A, B ou $AB^{-1}A$.

Cherchons les constantes γ_1 et γ_2. De même qu'au point 2, on démontre qu'on a l'égalité $\gamma_1 + \gamma_2 = 2$. Aussi dans les formules (6) donnant les paramètres d'itération τ_k, on a $\tau_0 = 2/(\gamma_1 + \gamma_2) = 1$, $\rho_0 = (\gamma_2 - \gamma_1)/(\gamma_2 + \gamma_1) = 1 - \gamma_1$.

Apprécions γ_1. A partir du lemme 14, § 2, ch. V, on tire que pour toute fonction de maille $\overset{\circ}{y} \in \overset{\circ}{H}$ on a l'inégalité

$$(b\overset{\circ}{y}, \overset{\circ}{y})_{\omega_\alpha} \leqslant \varkappa_\alpha \left[(a_\alpha \overset{\circ}{y}{}^2_{\bar{x}_\alpha}, 1)_{\omega_\alpha^+} + \frac{1}{2} (d\overset{\circ}{y}, \overset{\circ}{y})_{\omega_\alpha} \right], \quad \alpha = 1, 2, \tag{33}$$

où $\varkappa_\alpha = \varkappa_\alpha(x_\beta) = \max\limits_{x_\alpha \in \omega_\alpha} v^\alpha(x)$, tandis que $v^\alpha(x)$ est la solution du problème aux limites triponctuel suivant:

$$\begin{gathered}(a_\alpha v^\alpha_{\bar{x}_\alpha})_{x_\alpha} - \frac{1}{2} dv^\alpha = -b(x), \quad h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ v^\alpha(x) = 0, \quad x_\alpha = 0,\ l_\alpha \quad h_\beta \leqslant x_\beta \leqslant l_\beta - h_\beta, \quad \beta = 3 - \alpha, \quad \alpha = 1, 2.\end{gathered} \tag{34}$$

A partir des inégalités (33), en divisant par $\varkappa_\alpha$ et en sommant ensuite en ω_β, il vient

$$\left(\frac{b}{\varkappa_\alpha} \overset{\circ}{y}, \overset{\circ}{y} \right) \leqslant (a_\alpha \overset{\circ}{y}{}^2_{\bar{x}_\alpha}, 1) + \frac{1}{2} (d\overset{\circ}{y}, \overset{\circ}{y}), \quad \alpha = 1, 2.$$

En additionnant ces inégalités, compte tenu de (29), on obtient

$$\left(\sum_{\alpha=1}^{2} \frac{1}{\varkappa_\alpha} b\overset{\circ}{y}, \overset{\circ}{y} \right) \leqslant \sum_{\alpha=1}^{2} (a_\alpha \overset{\circ}{y}{}^2_{\bar{x}_\alpha}, 1) + (d\overset{\circ}{y}, \overset{\circ}{y}) = (Ay, y).$$

Par conséquent, on peut prendre en guise de γ_1

$$\gamma_1 = \min_{x \in \omega} \sum_{\alpha=1}^{2} \frac{1}{\varkappa_\alpha} = \min_{x_2 \in \omega_2} \frac{1}{\varkappa_1(x_2)} + \min_{x_1 \in \omega_1} \frac{1}{\varkappa_2(x_1)}. \tag{35}$$

Finalement, pour trouver γ_1, il faut résoudre l'équation (34), obtenir $\varkappa_\alpha(x_\beta)$ et à l'aide de la formule (35) calculer γ_1. La constante γ_2 s'obtient suivant la formule $\gamma_2 = 2 - \gamma_1$.

Cherchons l'estimation pour le nombre d'itérations de la méthode implicite de Tchébychev étudiée. De la théorie des schémas aux différences il s'ensuit que le schéma aux différences (34) est stable à droite en métrique régulière, c'est-à-dire qu'il existe une telle constante M, indépendante des pas du maillage h_1 et h_2, vérifiant pour la solution de l'équation (34) l'estimation

$$\max_{x_\alpha \in \omega_\alpha} v^\alpha(x) \leqslant M (b^2, 1)^{1/2}_{\omega_\alpha}.$$

Comme $b(x) = O\left(\frac{1}{h^2}\right)$, $h = \min\limits_{\alpha} h_\alpha$ pour $x \in \omega$, il s'ensuit que

$$\varkappa_\alpha = \max_{x_\alpha \in \omega_\alpha} v^\alpha(0) = O\left(\frac{1}{h^2}\right),$$

et, par suite, $\gamma_1 = O(h^2)$ et $\gamma_2 = O(1)$. Donc $\xi = \gamma_1/\gamma_2 = O(h^2)$, tandis que pour le nombre d'itérations on a une estimation asymptotique en h, identique à celle de la méthode explicite

$$n_0(\varepsilon) = O\left(\frac{1}{h} \ln \frac{2}{\varepsilon}\right).$$

En quoi consiste l'avantage de la méthode itérative implicite par rapport à la méthode explicite de Tchébychev étudiée auparavant? La réponse à cette question est fournie par le théorème suivant qu'on donnera sans démonstration.

Théorème 7 *). *Pour le schéma itératif* (5)-(7) *à opérateur A correspondant au schéma discret* (27), *le meilleur dans la classe d'opérateurs diagonaux B (c'est-à-dire pour lequel le quotient ξ est maximal) est l'opérateur défini par la formule* (32).

Il découle du théorème 7 que si en guise d'opérateur B on choisit la partie diagonale de l'opérateur A, le quotient $\xi = \gamma_1/\gamma_2$ des constantes de l'équivalence énergétique γ_1 et γ_2 des opérateurs A et B sera maximal et, partant, le nombre d'itérations n minimal.

Illustrons les avantages de la méthode implicite sur l'exemple modèle suivant. Soit le schéma aux différences (27) donné sur un maillage carré dans un carré unitaire $h_1 = h_2 = h = 1/N$, $l_1 = l_2 = 1$.

Les coefficients $a_1(x)$, $a_2(x)$ et $d(x)$ seront choisis de la façon suivante:

$$a_1(x) = 1 + c\,[(x_1 - 0{,}5)^2 + (x_2 - 0{,}5)^2],$$
$$a_2(x) = 1 + c\,[0{,}5 - (x_1 - 0{,}5)^2 - (x_2 - 0{,}5)^2],$$
$$d(x) \equiv 0,\ c > 0.$$

En outre, dans les inégalités (28) on a $c_1 = 1$, $c_2 = 1 + 0{,}5c$, $d_1 = d_2 = 0$. En variant le paramètre c, on obtient les coefficients du schéma aux différences (27) aux caractéristiques extrémales différentes.

On a montré que dans le cas de la méthode explicite le nombre d'itérations dépendait du quotient c_2/c_1. Pour la méthode implicite le nombre d'itérations est fonction non pas des valeurs maximale et minimale des coefficients $a_\alpha(x)$, mais de certaines caractéristiques intégrales de ces coefficients.

*) Ce théorème est un cas particulier d'un théorème plus général démontré dans: G. Forsythe, E. G. Straus, *On best conditioned matrices*, Proc. Amer. Math. Soc. 6 (1955), 340-345.

On a présenté au tableau 7 le nombre d'itérations associées aux méthodes explicite et implicite en fonction du quotient c_2/c_1 et du nombre de nœuds N suivant une direction. Les calculs sont effectués pour $\varepsilon = 10^{-4}$. C'est pour le cas où le paramètre $c = 0$, c'est-à-dire quand $a_\alpha(x) \equiv 1$, et que l'on étudie l'équation de Poisson, le nombre d'itérations des méthodes explicite et implicite est le même et est donné au point 1.

Tableau 7

$\frac{c_2}{c_1}$	$N = 32$		$N = 64$		$N = 128$	
	implicite	explicite	implicite	explicite	implicite	explicite
2	123	143	246	286	494	571
8	149	286	305	571	616	1142
32	175	571	365	1142	749	2283
128	192	1141	409	2283	856	4565
512	202	2281	436	4565	926	9130

Il s'ensuit du tableau que dans l'exemple étudié le nombre d'itérations de la méthode implicite est beaucoup inférieur à celui de la méthode explicite. La dépendance du nombre d'itérations du quotient c_2/c_1 est moindre pour la méthode implicite que pour la méthode explicite.

En guise de conclusion, donnons les formules de calcul pour la méthode implicite de Tchébychev:

$$y_{k+1}(i, j) = (1-\tau_{k+1})\, y_k(i, j) +$$
$$+\frac{\tau_{k+1}}{b(i, j)}\left\{\frac{1}{h_1^2}\left[a_1(i+1, j)\, y_k(i+1, j) + a_1(i, j)\, y_k(i-1, j)\right] +\right.$$
$$\left.+\frac{1}{h_2^2}\left[a_2(i, j+1)\, y_k(i, j+1) + a_2(i, j)\, y_k(i, j+1)\right] + \varphi(i, j)\right\},$$
$$1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant j \leqslant N_2 - 1,$$

où $b(i, j)$ est défini dans (32), tandis que l'approximation initiale y_0 est une fonction de maille arbitraire sur ω qui prend sur la frontière γ les valeurs données: $y_0(x) = g(x)$, $x \in \gamma$.

La comparaison des formules de calcul pour les méthodes explicite et implicite montre que le nombre d'opérations arithmétiques exigé pour le calcul de y_{k+1} sur la base de y_k fixé est pratiquement le même pour les deux méthodes. Etant donné que pour la méthode implicite le nombre d'itérations est sensiblement plus faible que pour la méthode explicite, il est naturel que la préférence soit accordée à la méthode implicite.

4. Problème discret de Dirichlet pour l'équation elliptique à dérivée mixte. Il s'agit de résoudre le *problème de Dirichlet pour l'équation elliptique à dérivées mixtes* dans le rectangle $\overline{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha, \alpha = 1, 2\}$ à frontière Γ

$$Lu = \sum_{\alpha, \beta=1}^{2} \frac{\partial}{\partial x_\alpha}\left(k_{\alpha\beta}(x)\frac{\partial u}{\partial x_\beta}\right) = -f(x), \quad x \in G,$$

$$u(x) = g(x), \quad x \in \Gamma.$$

On suppose que sont vérifiées les conditions de symétrie

$$k_{12}(x) = k_{21}(x), \quad x \in \overline{G}, \tag{36}$$

et d'ellipticité

$$c_1 \sum_{\alpha=1}^{2} \xi_\alpha^2 \leqslant \sum_{\alpha, \beta=1}^{2} k_{\alpha\beta}\xi_{\alpha}\xi_\beta \leqslant c_2 \sum_{\alpha=1}^{2} \xi_\alpha^2, \quad c_1 > 0, \tag{37}$$

où $\xi = (\xi_1, \xi_2)$ est un vecteur quelconque.

Sur un maillage rectangulaire

$$\overline{\omega} = \{x_{ij} = (ih_1, jh_2) \in \overline{G}, \ 0 \leqslant i \leqslant N_1, \ 0 \leqslant j \leqslant N_2,$$
$$h_\alpha = l_\alpha / N_\alpha, \ \alpha = 1, 2\}$$

au problème différentiel correspond le problème discret de Dirichlet

$$\Lambda y = 0,5 \sum_{\alpha, \beta=1}^{2} [(k_{\alpha\beta} y_{\bar{x}_\beta})_{x_\alpha} + (k_{\alpha\beta} y_{x_\beta})_{\bar{x}_\alpha}] = -\varphi(x), \quad x \in \omega, \tag{38}$$

$$y(x) = g(x), \quad x \in \gamma,$$

où γ est la frontière du maillage $\overline{\omega}$.

Au § 2, ch. V, on a montré que le problème de différences (38) se réduisait à l'équation opératorielle (1) de la façon banale: $Ay = \Lambda \mathring{y}$, où $y \in H$, $\mathring{y} \in \mathring{H}$ et $\mathring{y}(x) = y(x)$ pour $x \in \omega$. $\mathring{H}$ est ici un ensemble de fonctions de mailles données sur $\overline{\omega}$ et s'annulant sur γ, tandis que H est l'espace de fonctions de mailles données sur ω avec produit scalaire

$$(u, v) = \sum_{x \in \omega} u(x) v(x) h_1 h_2.$$

On y a également montré qu'avec la satisfaction de la condition (36) l'opérateur construit A est autoadjoint dans H et, si les conditions (37) sont remplies, possède des bornes γ_1 et γ_2 égales à

$$\gamma_1 = c_1 \sum_{\alpha=1}^{2} \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2 l_\alpha}, \quad \gamma_2 = c_2 \sum_{\alpha=1}^{2} \frac{4}{h_\alpha^2} \cos^2 \frac{\pi h_\alpha}{2 l_\alpha}, \tag{39}$$

c'est-à-dire

$$\gamma_1 E \leqslant A \leqslant \gamma_2 E. \tag{40}$$

Pour résoudre de façon approchée l'équation (1) correspondant au schéma discret (38), rapportons-nous à la méthode explicite de Tchébychev (5)-(7) ($B = E$). Vu que les opérateurs A et B sont autoadjoints et définis positifs dans H, l'information à priori prend la forme de constantes γ_1 et γ_2 dans les inégalités (40) et la méthode converge dans H_D, où $D = A$, B ou $AB^{-1}A$.

A partir de (39) on obtient

$$\gamma_1 = O(c_1), \quad \gamma_2 = O\left(\frac{c_2}{h^2}\right), \quad \xi = \frac{\gamma_1}{\gamma_2} = O\left(\frac{c_1}{c_2} h^2\right), \quad h^2 = h_1^2 + h_2^2.$$

Donc pour l'exemple envisagé l'estimation asymptotique en h du nombre d'itérations $n_0(\varepsilon)$ prend la forme

$$n_0(\varepsilon) = O\left(\sqrt{\frac{c_2}{c_1}} \frac{1}{h} \ln \frac{2}{\varepsilon}\right).$$

Dans le cas particulier où $\overline{G}$ est un carré de côté l et le maillage $\overline{\omega}$ est également carré ($h_1 = h_2 = h = l/N$), il vient

$$\gamma_1 = \frac{8c_1}{h^2} \sin^2 \frac{\pi h}{2l}, \quad \gamma_2 = \frac{8c_2}{h^2} \cos^2 \frac{\pi h}{2l}, \quad \xi = \frac{c_1}{c_2} \operatorname{tg}^2 \frac{\pi h}{2l},$$

$$n \geqslant n_0(\varepsilon) \approx \sqrt{\frac{c_2}{c_1}} \frac{l}{\pi h} \ln \frac{2}{\varepsilon} = 0{,}32 \sqrt{\frac{c_2}{c_1}} N \ln \frac{2}{\varepsilon},$$

c'est-à-dire que le nombre d'itérations est également proportionnel au nombre de nœuds N suivant une direction, comme au cas d'équation sans dérivées mixtes.

Sur ce, on achèvera l'étude d'exemples d'application des méthodes itératives à deux couches à la résolution d'équations elliptiques. Des exemples plus compliqués seront passés en revue dans le chapitre XIV.

CHAPITRE VII

MÉTHODES ITÉRATIVES À TROIS COUCHES

On étudie dans ce chapitre les méthodes itératives à trois couches permettant de résoudre l'équation opératorielle $Au = f$. Les paramètres d'itération sont choisis avec la prise en compte de l'information à priori sur les opérateurs du schéma. Au § 1 on apprécie la vitesse de convergence du schéma à trois couches du type standard. Les §§ 2, 3 traitent de la méthode semi-itérative de Tchébychev et de la méthode de stationnarisation à trois couches. Le § 4 est consacré à l'étude de la stabilité des méthodes à deux et à trois couches par rapport aux perturbations des données à priori.

§ 1. Appréciation de la vitesse de convergence

1. Famille de départ des schémas itératifs. Au ch. VI, pour trouver la solution approchée de l'équation opératorielle linéaire

$$Au = f \tag{1}$$

à opérateur A non dégénéré agissant dans l'espace hilbertien réel H, on a construit des schémas itératifs à deux couches. Dans ces méthodes le schéma à deux couches relie deux approximations itératives y_{k+1} et y_k.

Le présent chapitre sera consacré à l'étude des schémas itératifs à trois couches. Le schéma itératif à trois couches pour l'équation (1) relie trois approximations itératives y_{k+1}, y_k et y_{k-1}, de sorte que y_{k+1} est défini au moyen de y_k et y_{k-1}. Pour la mise en œuvre du schéma à trois couches il est nécessaire de fixer deux approximations initiales y_0 et y_1. Généralement avec un y_0 arbitraire on obtient l'approximation y_1 suivant le schéma à deux couches.

Limitons-nous à l'étude des schémas à trois couches du type standard. Le *schéma itératif à trois couches standard* de nature implicite est de la forme

$$\begin{gathered} By_{k+1} = \alpha_{k+1}(B - \tau_{k+1}A)\, y_k + (1 - \alpha_{k+1})\, By_{k-1} + \alpha_{k+1}\tau_{k+1}f, \\ By_1 = (B - \tau_1 A)\, y_0 + \tau_1 f, \quad k = 1, 2, \ldots, \quad y_0 \in H, \end{gathered} \tag{2}$$

où y_0 est une approximation initiale quelconque, B un opérateur linéaire non dégénéré, agissant dans H, α_k et τ_k des paramètres d'itération. Les formules (2) définissent la famille de départ des schémas itératifs à trois couches.

22—01162

L'obtention de la nouvelle approximation y_{k+1} peut être exprimée de la façon suivante. Soit $\bar{y}$ l'approximation itérative intermédiaire qu'on obtient suivant le schéma implicite à deux couches

$$B\frac{\bar{y}-y_k}{\tau_{k+1}}+Ay_k=f.$$

Il s'ensuit alors de (2) que y_{k+1} est une combinaison linéaire des approximations $\bar{y}$ et y_{k-1}

$$y_{k+1}=\alpha_{k+1}\bar{y}+(1-\alpha_{k+1})\,y_{k-1}.$$

L'approximation y_{k+1} est donc une extrapolation linéaire des approximations $\bar{y}$ et y_{k-1}.

Si l'on pose dans (2) $\alpha_k \equiv 1$, le schéma à trois couches (2) passe au schéma à deux couches dont la convergence a été étudiée au ch. VI. Par conséquent, l'introduction des paramètres d'itération α_k permet d'espérer que la convergence du schéma (2) sera non inférieure à celle du schéma à deux couches.

Notons qu'à la différence du schéma itératif à deux couches la mise en œuvre du schéma à trois couches oblige à mémoriser non pas une mais deux approximations itératives y_k et y_{k-1}.

2. Appréciation de la norme d'erreur. Abordons maintenant l'étude de la convergence du schéma à trois couches (2) dans l'espace énergétique H_D engendré par l'opérateur D autoadjoint et défini positif dans H. A cette fin voyons le comportement dans H_D de la norme d'erreur $z_k = y_k - u$ pour $k \to \infty$.

En portant $y_k = z_k + u$ pour $k = 0, 1, \ldots$ dans (2) et en tenant compte de l'équation (1), on obtient l'équation pour l'erreur z_k:

$$Bz_{k+1} = \alpha_{k+1}(B-\tau_{k+1}A)\,z_k + (1-\alpha_{k+1})\,Bz_{k-1}, \quad k = 1, 2, \ldots,$$
$$Bz_1 = (B-\tau_1 A)\,z_0, \quad z_0 = y_0 - u.$$

Résolvons cette équation par rapport à z_{k+1} et, en posant $z_k = D^{-1/2}x_k$, passons à l'équation pour l'erreur équivalente x_k. L'équation pour x_k prendra la forme suivante:

$$\begin{aligned} x_{k+1} &= \alpha_{k+1}S_{k+1}x_k + (1-\alpha_{k+1})\,x_{k-1}, \quad k = 1, 2, \ldots, \\ x_1 &= S_1x_0, \quad S_k = E-\tau_kC, \end{aligned} \tag{3}$$

où $C = D^{1/2}B^{-1}AD^{-1/2}$.

En raison de la substitution réalisée $z_k = D^{-1/2}x_k$ l'égalité $\|x_k\| = \|z_k\|_D$ doit se vérifier et, partant, le schéma (2) convergera dans H_D si $\|x\|_k \to 0$ pour $k \to \infty$.

Etudions la conduite de la norme x_k dans H pour $k \to \infty$. A cette fin cherchons la forme explicite de la solution de l'équation (9).

En utilisant les formules (9), on obtient successivement

$$x_1 = (E - \tau_1 C)\, x_0 = P_1(C)\, x_0,$$

$$x_2 = \alpha_2 (E - \tau_2 C)\, x_1 + (1 - \alpha_2)\, x_0 = [\alpha_2 (E - \tau_2 C)\, P_1(C) + (1 - \alpha_2)\, E]\, x_0 = P_2(C)\, x_0,$$

. . . .

$$x_{k+1} = \alpha_{k+1} (E - \tau_{k+1} C)\, P_k(C)\, x_0 + (1 - \alpha_{k+1})\, P_{k-1}(C)\, x_0 = P_{k+1}(C)\, x_0, \text{ etc.}$$

Par conséquent, la solution de l'équation (9) pour tout k prend la forme

$$x_k = P_k(C)\, x_0, \quad k = 0, 1, \ldots, \qquad (4)$$

où $P_k(C)$ est un polynôme opératoriel de degré k relativement à l'opérateur C. En vertu de l'arbitraire de x_0, le polynôme algébrique correspondant $P_k(t)$ satisfait aux relations de récurrence suivantes:

$$P_{k+1}(t) = \alpha_{k+1}(1 - \tau_{k+1} t)\, P_k(t) + (1 - \alpha_{k+1})\, P_{k-1}(t),$$
$$P_1(t) = 1 - \tau_1 t, \quad P_0(t) \equiv 1, \quad k = 1, 2, \ldots \qquad (5)$$

Il s'ensuit de (5) que le polynôme $P_k(t)$ remplit la condition de normalisation de $P_k(0) = 1$ pour tout k.

Apprécions maintenant la norme x_k. De (4), il vient

$$\| x_k \| = \| P_k(C)\, x_0 \| \leqslant \| P_k(C) \| \, \| x_0 \|, \quad k = 0, 1, \ldots$$

ou, en vertu de la substitution effectuée $z_k = D^{-1/2} x_k$,

$$\| z_k \|_D \leqslant \| P_k(C) \| \, \| z_0 \|_D. \qquad (6)$$

Bref, l'estimation de la norme de l'erreur z_k est obtenue. De (6) il s'ensuit que la méthode aura la plus grande vitesse de convergence au cas où la norme du polynôme $P_k(C)$ tendra le plus rapidement possible vers zéro pour $k \to \infty$. Etant donné que le polynôme $P_k(C)$ est une fonction de paramètres d'itération $\tau_1, \tau_2, \ldots, \tau_k$ et $\alpha_2, \alpha_3, \ldots, \alpha_k$, il faut choisir ces paramètres sur la base de la condition du minimum de la norme du polynôme opératoriel $P_k(C)$. Autrement dit, il faut construire un polynôme de degré k normé par la condition $P_k(0) = E$, dont la norme est minimale.

Au ch. VI, lors de l'étude de la méthode de Tchébychev à deux couches, ce problème a été résolu dans l'hypothèse que l'opérateur $DB^{-1}A$ est autoadjoint dans H et que sont données les constantes de l'équivalence énergétique γ_1 et γ_2 des opérateurs autoadjoints D et $DB^{-1}A$:

$$\gamma_1 D \leqslant DB^{-1}A \leqslant \gamma_2 D, \quad \gamma_1 > 0, \quad DB^{-1}A = (DB^{-1}A)^*. \qquad (7)$$

Lors de la construction des méthodes itératives à trois couches on n'étudiera que le cas d'autoconjugaison, c'est-à-dire qu'on posera remplies les conditions (7).

Les conditions (7) étant remplies, formons l'opérateur optimal $P_k(C)$ et obtenons l'estimation à priori pour l'erreur z_k.

Etant donné que $C = D^{1/2}B^{-1}AD^{-1/2} = D^{-1/2}(DB^{-1}A)D^{-1/2}$, il s'ensuit de (7) que l'opérateur C est autoadjoint dans H, tandis que γ_1 et γ_2 sont ses bornes:

$$\gamma_1 E \leqslant C \leqslant \gamma_2 E, \quad \gamma_1 > 0, \quad C = C^*. \tag{8}$$

Dans ce cas, en vertu de (8), pour la norme de l'opérateur $P_k(C)$ se vérifie l'estimation

$$\| P_k(C) \| \leqslant \max_{\gamma_1 \leqslant t \leqslant \gamma_2} |P_k(t)|.$$

Par conséquent, le polynôme optimal $P_k(t)$ se dégage de la condition suivante: le maximum du module de ce polynôme sur le segment $[\gamma_1, \gamma_2]$ est minimal. Du § 2, ch. VI, une fois posée la condition de normalisation du polynôme $P_k(0) = 1$, il s'ensuit que le polynôme cherché a l'aspect

$$P_k(t) = q_k T_k\left(\frac{1-\tau_0 t}{\rho_0}\right), \quad q_k = \frac{1}{T_k\left(\frac{1}{\rho_0}\right)}, \tag{9}$$

où $T_k(x)$ est le polynôme de Tchébychev de première espèce et de degré k:

$$T_k(x) = \begin{cases} \cos(k \arccos x), & |x| \leqslant 1, \\ \operatorname{ch}(k \operatorname{Arch} x), & |x| \geqslant 1, \end{cases}$$

$$\tau_0 = \frac{2}{\gamma_1+\gamma_2}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad q_k = \frac{2\rho_1^k}{1+\rho_1^{2k}}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

On a dans ce cas l'estimation

$$\| P_k(C) \| \leqslant \max_{\gamma_1 \leqslant t \leqslant \gamma_2} |P_k(t)| = q_k, \quad k = 0, 1, \ldots$$

En portant cette estimation dans (6), il vient

$$\| z_k \|_D \leqslant q_k \| z_0 \|_D.$$

Ainsi donc la vitesse de convergence de la méthode itérative à trois couches (2), dont les paramètres d'itération τ_k et α_k sont choisis sur la base de la condition du minimum de la norme de l'opérateur résolvant, est égale à la vitesse de convergence de la méthode itérative de Tchébychev à deux couches.

Les formules (9) fournissent la solution du problème de la plus rapide construction de la méthode itérative à trois couches. On obtiendra au § 2 les formules des paramètres d'itération τ_k et α_k de cette méthode appelée *méthode semi-itérative de Tchébychev*.

§ 2. Méthode semi-itérative de Tchébychev

1. Formules des paramètres d'itération. Cherchons maintenant les formules pour les paramètres d'itération α_k et τ_k de la *méthode semi-itérative de Tchébychev*. Au § 1, en utilisant le schéma itératif à trois couches de la méthode

$$By_{k+1} = \alpha_{k+1}(B - \tau_{k+1}A)\, y_k + (1 - \alpha_{k+1})\, By_{k-1} + \alpha_{k+1}\tau_{k+1}f, \tag{1}$$
$$By_1 = (B - \tau_1 A)\, y_0 + \tau_1 f, \quad k = 1, 2, \ldots, \quad y_0 \in H,$$

on a obtenu l'équation pour l'erreur équivalente

$$x_{k+1} = \alpha_{k+1}(E - \tau_{k+1}C)\, x_k + (1 - \alpha_{k+1})\, x_{k-1}, \quad k = 1, 2, \ldots, \tag{2}$$
$$x_1 = (E - \tau_1 C)\, x_0.$$

On a montré que pour tout k la solution de cette équation est de la forme

$$x_k = P_k(C)\, x_0, \quad k = 0, 1, \ldots, \tag{3}$$

tandis que le polynôme optimal $P_k(C)$ est déterminé par les formules

$$P_k(t) = q_k T_k\left(\frac{1 - \tau_0 t}{\rho_0}\right), \quad q_k = \frac{1}{T_k\left(\frac{1}{\rho_0}\right)} = \frac{2\rho_1^k}{1 + \rho_1^{2k}}. \tag{4}$$

Pour obtenir les formules des paramètres d'itération α_k et τ_k cherchons les relations de récurrence que vérifie le polynôme $P_k(t)$.

Il est connu que pour tout x les polynômes de Tchébychev de première espèce $T_k(x)$ vérifient les relations de récurrence suivantes (voir § 4, ch. I):

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x), \quad k = 1, 2, \ldots, \qquad T_1(x) = x, \quad T_0(x) \equiv 1. \tag{5}$$

En utilisant (4) et (5), il vient

$$\frac{P_{k+1}(t)}{q_{k+1}} = 2\left(\frac{1 - \tau_0 t}{\rho_0}\right)\frac{P_k(t)}{q_k} - \frac{P_{k-1}(t)}{q_{k-1}}, \quad k = 1, 2, \ldots, \tag{6}$$
$$P_1(t)/q_1 = (1 - \tau_0 t)/\rho_0, \quad P_0(t)/q_0 \equiv 1. \tag{7}$$

De la définition (4) et des relations (5) il découle que

$$1/q_{k+1} = 2/(\rho_0 q_k) - 1/q_{k-1}, \quad q_1 = \rho_0, \quad q_0 = 1. \tag{8}$$

De là il vient

$$q_{k+1}/q_{k-1} = 2q_{k+1}/(\rho_0 q_k) - 1, \quad k = 1, 2, \ldots \tag{9}$$

En portant (8), (9) dans (6) et (7), on obtient les formules de récurrence pour le polynôme $P_k(t)$:

$$P_{k+1}(t) = \frac{2}{\rho_0}\frac{q_{k+1}}{q_k}(1 - \tau_0 t)\, P_k(t) + \left(1 - \frac{2}{\rho_0}\frac{q_{k+1}}{q_k}\right) P_{k-1}(t),$$
$$P_1(t) = 1 - \tau_0 t, \quad P_0(t) \equiv 1, \quad k = 1, 2, \ldots$$

De là et à partir de (3) s'ensuivent les relations de récurrence pour x_k

$$x_{k+1} = \frac{2}{\rho_0}\frac{q_{k+1}}{q_k}(E-\tau_0 C)x_k + \left(1-\frac{2}{\rho_0}\frac{q_{k+1}}{q_k}\right)x_{k-1}, \quad k=1, 2, \ldots,$$
$$x_1 = (E-\tau_0 C)x_0.$$

En comparant ces formules avec (2), il vient

$$\alpha_{k+1} = 2q_{k+1}/(\rho_0 q_k), \quad \tau_k \equiv \tau_0 = 2/(\gamma_1+\gamma_2), \quad k = 1, 2, \ldots \quad (10)$$

On a ainsi obtenu les formules des paramètres d'itération τ_k et α_k. Transformons la formule pour les paramètres α_k. Pour cela calculons, en utilisant (8), l'expression

$$4\frac{\alpha_{k+1}-1}{\alpha_k\alpha_{k+1}} = \frac{4}{\alpha_k}\left(1-\frac{1}{\alpha_{k+1}}\right) = \frac{4}{\alpha_k}\left(1-\frac{\rho_0}{2}\frac{q_k}{q_{k+1}}\right) =$$
$$= \frac{4}{\alpha_k}\left[1-\frac{\rho_0}{2}\left(\frac{2}{\rho_0}-\frac{q_k}{q_{k-1}}\right)\right] = \frac{2\rho_0}{\alpha_k}\frac{q_k}{q_{k-1}} = \rho_0^2.$$

De là on obtient $\alpha_{k+1} = 4/(4-\rho_0^2\alpha_k)$, $k = 1, 2, \ldots$ En posant dans (10) $k = 0$ et tenant compte de (8), on trouve que $\alpha_1 = 2$.

On a ainsi démontré le

Théorème 1. *Supposons que sont satisfaites les conditions*

$$\gamma_1 D \leqslant DB^{-1}A \leqslant \gamma_2 D, \quad \gamma_1 > 0, \quad DB^{-1}A = (DB^{-1}A)^*.$$

La méthode semi-itérative de Tchébychev (2) *à paramètres d'itération*

$$\tau_k \equiv 2/(\gamma_1+\gamma_2), \quad \alpha_{k+1} = 4/(4-\rho_0^2\alpha_k), \quad k = 1, 2, \ldots,$$
$$\alpha_1 = 2, \quad (11)$$

converge dans H_D et pour l'erreur z_k se vérifie l'estimation

$$\| z_k \|_D \leqslant q_k \| z_0 \|_D.$$

Pour le nombre d'itérations n on a l'estimation $n \geqslant n_0(\varepsilon)$, où

$$n_0(\varepsilon) = \frac{\ln 0{,}5\varepsilon}{\ln \rho_1}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad q_k = \frac{2\rho_1^k}{1+\rho_1^{2k}}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

Remarque. La confrontation de la méthode semi-itérative de Tchébychev et de la méthode de Tchébychev à deux couches montre qu'on a pour ces méthodes une même estimation $\| z_n \|_D \leqslant q_n \| z_0 \|_D$, si n itérations sont mises en œuvre. Toutefois, pour la méthode à deux couches cette estimation n'est vraie qu'après l'exécution de toutes les itérations, tandis que pour la méthode à trois couches elle se vérifie pour toutes itérations intermédiaires. A la différence de la méthode à deux couches, dans la méthode à trois couches les normes d'erreur diminuent de façon monotone sur les itérations intermédiaires, ce qui garantit la stabilité des calculs de la méthode à trois couches.

2. Exemples de choix de l'opérateur D. Donnons maintenant des exemples de choix de l'opérateur D et l'exigence imposée aux opérateurs A et B vérifiant les conditions du théorème 1.

Au point 3, § 2, ch. VI, on a passé en revue quelques cas de choix de l'opérateur D en fonction des propriétés des opérateurs A et B. Rappelons ces résultats.

1) Si les opérateurs A et B sont autoadjoints et définis positifs dans H, on peut choisir en guise de D l'un des opérateurs suivants: A, B ou $AB^{-1}A$. L'information à priori peut alors être présentée sous forme

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \quad \gamma_1 > 0. \tag{12}$$

2) Si les opérateurs A et B sont autoadjoints, définis positifs et permutables $A = A^* > 0$, $B = B^* > 0$, $AB = BA$, on peut alors choisir en guise de D l'opérateur A^*A. Dans ce cas l'information à priori prend la forme des inégalités (12).

3) Si A et B sont des opérateurs non dégénérés satisfaisant à la condition $B^*A = A^*B$, en qualité de D on peut également recourir à l'opérateur A^*A. L'information à priori se présente alors sous forme des inégalités

$$\gamma_1 (Bx, Bx) \leqslant (Ax, Bx) \leqslant \gamma_2 (Bx, Bx), \quad \gamma_1 > 0.$$

Ces hypothèses étant vérifiées, on peut appliquer à la méthode semi-itérative de Tchébychev à trois couches le théorème 1.

3. L'algorithme de la méthode. Etudions le problème de la mise en œuvre du schéma à trois couches (1). L'algorithme de la méthode peut être décrit de la façon suivante:

1) en fonction de la valeur du paramètre α_k et des approximations données y_{k-1} et y_k on trouve α_{k+1} et τ_{k+1}, en appliquant les formules (11) et l'on calcule

$$\varphi = B(\alpha_{k+1} y_k + (1 - \alpha_{k+1}) y_{k-1}) - \alpha_{k+1}\tau_{k+1}(Ay_k - f).$$

φ une fois calculé peut être placé à l'endroit de y_{k-1} qui, pour les itérations suivantes, est déjà inutile;

2) pour obtenir la nouvelle approximation y_{k+1} on résout l'équation $By_{k+1} = \varphi$. L'approximation y_1 se déduit de l'équation $By_1 = \varphi$, où $\varphi = By_0 - \tau_1 (Ay_0 - f)$. Cet algorithme de résolution peut être recommandé dans le cas où il est nécessaire d'économiser la mémoire de l'ordinateur.

Si le calcul de la valeur de l'opérateur B se solde par un grand nombre d'opérations arithmétiques et il n'est pas nécessaire d'économiser la mémoire, il est conseillé de recourir à l'algorithme suivant:

1) en fonction de y_k fixé on calcule le résidu $r_k = Ay_k - f$;

2) on résout l'équation de correction w_k: $Bw_k = r_k$;

3) en fonction de α_k donné on calcule α_{k+1} suivant la formule (11), tandis que la nouvelle approximation s'obtient par la formule

$$y_{k+1} = \alpha_{k+1} y_k + (1 - \alpha_{k+1})\, y_{k-1} - \alpha_{k+1}\tau_{k+1} w_k,$$

où τ_{k+1} est déterminé d'après la formule (11).

L'algorithme décrit ne contient pas de calcul de la valeur de l'opérateur B, mais exige une mémoire complémentaire pour la mémorisation de r_k et w_k.

§ 3. Méthode de stationnarisation à trois couches

1. Choix des paramètres d'itération. Revonons maintenant aux formules des paramètres d'itération α_k et τ_k de la méthode semi-itérative de Tchébychev. On a obtenu au § 1 les expressions suivantes pour α_{k+1} et τ_{k+1}:

$$\alpha_{k+1} = 2q_{k+1}/(\rho_0 q_k), \quad \tau_k \equiv \tau_0 = 2/(\gamma_1 + \gamma_2), \quad k = 1, 2, \ldots, \tag{1}$$

où

$$q_k = \frac{2\rho_1^k}{1+\rho_1^{2k}}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}. \tag{2}$$

La valeur du paramètre d'itération τ_k est indépendante du numéro d'itération k, tandis que le paramètre α_k varie à partir de $\alpha_1 = 2$. Cherchons la valeur limite pour α_k quand k tend vers l'infini. A partir de (1), (2), il vient

$$\alpha_{k+1} = 2\rho_1\,(1 + \rho_1^{2k})/(\rho_0\,(1 + \rho_1^{2k+2})).$$

Etant donné que $\rho < 1$ et $\rho_0 = q_1 = 2\rho_1/(1 + \rho_1^2)$, $\alpha = \lim\limits_{k\to\infty} \alpha_k = 1 + \rho_1^2$ et pour des k suffisamment grands, on a $\alpha_k \approx \alpha$. Aussi est-il naturel de procéder à l'étude de la *méthode itérative de stationnarisation à trois couches*

$$By_{k+1} = \alpha\,(B - \tau A)\, y_k + (1 - \alpha)\, By_{k-1} + \alpha\tau f, \quad k = 1, 2, \ldots, \tag{3}$$

$$By_1 = (B - \tau A)\, y_0 + \tau f, \quad y_0 \in H$$

à paramètres constants (stationnaires)

$$\alpha = 1 + \rho_1^2, \quad \tau = \tau_0 = \frac{2}{\gamma_1+\gamma_2}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}, \tag{4}$$

où γ_1 et γ_2 sont des constantes de l'équivalence énergétique des opérateurs autoadjoints D et $DB^{-1}A$:

$$\gamma_1 D \leqslant DB^{-1}A \leqslant \gamma_2 D, \quad \gamma_1 > 0, \quad DB^{-1}A = (DB^{-1}A)^*. \tag{5}$$

2. Appréciation de la vitesse de convergence. Pour obtenir l'estimation de la convergence de la méthode de stationnarisation à trois couches, passons de (3) au schéma pour une erreur équivalente $x_k =$

$= D^{1/2}z_k$:

$$x_{k+1} = \alpha (E - \tau C) x_k + (1 - \alpha) x_{k-1}, \quad k = 1, 2, \ldots,$$
$$x_1 = (E - \tau C) x_0, \quad C = D^{1/2}B^{-1}AD^{-1/2}.$$

Il s'ensuit de là que x_k s'exprime pour tout $k \geqslant 0$ au moyen de x_0 de la manière suivante:

$$x_k = P_k (C) x_0, \tag{6}$$

où le polynôme algébrique $P_k (t)$, correspondant à $P_k (C)$, se détermine au moyen des relations de récurrence

$$\begin{gathered} P_{k+1} (t) = \alpha (1 - \tau t) P_k (t) + (1 - \alpha) P_{k-1} (t), \quad k = 1, 2, \ldots, \\ P_1 (t) = 1 - \tau t, \quad P_0 (t) \equiv 1. \end{gathered} \tag{7}$$

De (6) on déduit l'estimation de la norme d'erreur z_k dans H_D:

$$\| z_k \|_D = \| x_k \| \leqslant \| P_k (C) \| \, \| x_0 \| = \| P_k (C) \| \, \| z_0 \|_D. \tag{8}$$

Il faut donc apprécier la norme du polynôme opératoriel $P_k (C)$ pour le cas où les paramètres α et τ sont choisis sur la base des formules (4). Il s'ensuit des conditions (5) que C est un opérateur autoadjoint dans H, tandis que γ_1 et γ_2 sont ses bornes, et, partant,

$$\| P_k (C) \| \leqslant \max_{\gamma_1 \leqslant t \leqslant \gamma_2} | P_k (t) |.$$

Apprécions le maximum du module du polynôme $P_k (t)$ sur le segment $[\gamma_1, \gamma_2]$. A cette fin exprimons le polynôme $P_k (t)$ au moyen du polynôme de Tchébychev. Il est plus commode d'étudier le polynôme $P_k (t)$ non pas sur le segment $[\gamma_1, \gamma_2]$ mais sur le segment standard $[-1, 1]$. En posant

$$t = \frac{1 - \rho_0 x}{\tau_0}, \quad \tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2},$$

représentons le segment $[\gamma_1, \gamma_2]$ sur le segment $[-1, 1]$. On a alors

$$P_k (t) = Q_k (x), \quad x \in [-1, 1],$$
$$\max_{\gamma_1 \leqslant t \leqslant \gamma_2} | P_k (t) | = \max_{|x| \leqslant 1} | Q_k (x) |.$$

Compte tenu du choix des paramètres α et τ en conformité de (4), on obtient de (7) les relations de récurrence suivantes pour les polynômes $Q_k (x)$:

$$Q_{k+1} (x) = 2\rho_1 x Q_k (x) - \rho_1^2 Q_{k-1} (x), \quad k = 1, 2, \ldots,$$
$$Q_1 (x) = \rho_0 x, \quad Q_0 (x) \equiv 1.$$

De là au moyen de la substitution

$$Q_k = \rho_1^k R_k (x) \tag{9}$$

on obtient sans peine la relation de récurrence standard

$$R_{k+1}(x) = 2xR_k(x) - R_{k-1}(x), \quad k = 1, 2, \ldots,$$
$$R_1(x) = \rho_0 x/\rho_1, \quad R_0(x) \equiv 1. \tag{10}$$

A cette relation répondent le polynôme de Tchébychev de première espèce $T_k(x)$ aux conditions initiales $T_k(x) = x$, $T_0(x) \equiv 1$ et le polynôme de Tchébychev de seconde espèce $U_k(x)$:

$$U_k(x) = \begin{cases} \dfrac{\sin((k+1)\arccos x)}{\sin(\arccos x)}, & |x| \leqslant 1, \\ \dfrac{\operatorname{sh}((k+1)\operatorname{Arch} x)}{\operatorname{sh}(\operatorname{Arch} x)}, & |x| \geqslant 1, \end{cases}$$

aux conditions initiales $U_1(x) = 2x$, $U_0(x) \equiv 1$. En utilisant les propriétés mentionnées des polynômes $T_k(x)$ et $U_k(x)$ et l'égalité $\rho_0 = q_1 = 2\rho_1/(1 + \rho_1^2)$, on obtient à partir de (10) l'expression du polynôme $R_k(x)$ en fonction des polynômes de Tchébychev

$$R_k(x) = \frac{2\rho_1^2}{1+\rho_1^2} T_k(x) + \frac{1-\rho_1^2}{1+\rho_1^2} U_k(x), \quad k \geqslant 0.$$

Ensuite, en utilisant les estimations connues

$$\max_{|x| \leqslant 1} |T_k(x)| = T_k(1) = 1,$$
$$\max_{|x| \leqslant 1} |U_k(x)| = U_k(1) = k+1,$$

on obtient

$$\max_{|x| \leqslant 1} |R_k(x)| = R_k(1) = 1 + k(1-\rho_1^2)/(1+\rho_1^2).$$

De là, compte tenu des substitutions faites plus haut, on tire l'estimation suivante de la norme du polynôme opératoriel $P_k(C)$:

$$\| P_k(C) \| \leqslant \rho_1^k (1 + k(1 - \rho_1^2)/(1 + \rho_1^2)). \tag{11}$$

En portant (11) dans (8), on obtient l'estimation pour la norme d'erreur z_k dans H_D:

$$\| z_k \|_D \leqslant \bar{q}_k \| z_0 \|_D, \quad \bar{q}_k = \rho_1^k (1 + k(1-\rho_1^2)/(1+\rho_1^2)),$$

de plus, $\bar{q}_k \to 0$ pour $k \to \infty$ et $\bar{q}_{k+1} < \bar{q}_k$. On a ainsi démontré le

Théorème 2. *La méthode itérative de stationnarisation à trois couches* (3)-(5) *converge dans* H_D *et pour l'erreur* z_k *se vérifie l'estimation*

$$\| z_k \|_D \leqslant \bar{q}_k \| z_0 \|_D, \quad \bar{q}_k = \rho_1^k (1 + k(1-\rho_1^2)/(1+\rho_1^2)).$$

Remarque. On peut montrer que $\lim\limits_{k\to\infty} q_k/\bar{q}_k = \lim\limits_{k\to\infty} \bar{q}_k/\rho_0^k = 0$, où q_k est défini dans le théorème 1. Aussi la méthode de station-

narisation à trois couches converge plus vite que la méthode itérative simple, mais moins vite que la méthode de Tchébychev à deux couches et la méthode semi-itérative de Tchébychev.

§ 4. Stabilité des méthodes à deux et à trois couches avec information à priori

1. Position du problème. Pour la résolution approchée de l'équation opératorielle $Au = f$ on a étudié au ch. VI la méthode itérative simple à deux couches et la méthode de Tchébychev et aux §§ 2, 3. ch. VII, on a construit la méthode semi-itérative de Tchébychev et la méthode itérative de stationnarisation à trois couches.

Rappelons que pour le calcul des paramètres d'itération on utilise dans ces méthodes une information à priori déterminée sur les opérateurs du schéma itératif. Au cas où l'opérateur $DB^{-1}A$ est autoadjoint, cette information prend la forme des constantes de l'équivalence énergétique γ_1 et γ_2 des opérateurs D et $DB^{-1}A$:

$$\gamma_1 (Dx, x) \leqslant (DB^{-1}Ax, x) \leqslant \gamma_2 (Dx, x), \quad \gamma_1 > 0. \tag{1}$$

En maintes occasions les constantes γ_1 et γ_2 peuvent être recherchées de façon précise, c'est-à-dire qu'il existe des éléments $x \in H$ pour lesquels il y a égalité dans (1). Dans d'autres cas pour obtenir γ_1 et γ_2 on recourt à des procédés auxiliaires et ces constantes sont établies de façon approchée.

L'utilisation d'une information à priori imprécise se solde par une diminution de la vitesse de convergence et même dans certains cas aboutit à la divergence de la méthode. L'objectif de ce paragraphe est d'élucider l'influence de l'information à priori imprécise sur la convergence des méthodes itératives mentionnées plus haut.

Bornons-nous à l'étude du cas d'autoconjugaison, c'est-à-dire admettons que l'opérateur $DB^{-1}A$ est autoadjoint dans H. Supposons que dans les inégalités (1) au lieu des valeurs précises γ_1 et γ_2 figurent des valeurs approchées $\tilde{\gamma}_1$ et $\tilde{\gamma}_2$. Abordons l'étude des méthodes à deux et trois couches dont les paramètres d'itération seront choisis d'après $\tilde{\gamma}_1$ et $\tilde{\gamma}_2$. Rappelons les formules donnant les paramètres d'itération.

Pour le schéma à deux couches

$$B\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \ldots, \tag{2}$$

les paramètres de la méthode itérative simple se déterminent à l'aide de la formule

$$\tau_k = \tilde{\tau}_0 = 2/(\tilde{\gamma}_1 + \tilde{\gamma}_2), \quad k = 1, 2, \ldots, \tag{3}$$

quant aux paramètres de la méthode de Tchébychev, ils s'obtiennent suivant la formule

$$\tau_k = \tilde{\tau}_0 / (1 + \tilde{\rho}_0 \mu_k), \quad \mu_k \in \mathfrak{M}_n^*, \quad k = 1, 2, \ldots, n,$$
$$\tilde{\rho}_0 = (1 - \tilde{\xi})/(1 + \tilde{\xi}), \quad \tilde{\xi} = \tilde{\gamma}_1/\tilde{\gamma}_2. \tag{4}$$

Pour le schéma itératif à trois couches

$$By_{k+1} = \alpha_{k+1} (B - \tau_{k+1} A)\, y_k + (1 - \alpha_{k+1})\, By_{k-1} + \alpha_{k+1} \tau_{k+1} f,$$
$$k = 1, 2, \ldots, \tag{5}$$
$$By_1 = (B - \tau_1 A)\, y_0 + \tau_1 f$$

les paramètres de la méthode semi-itérative de Tchébychev se déterminent suivant les formules

$$\tau_k \equiv \tilde{\tau}_0, \quad \alpha_{k+1} = 4/(4 - \tilde{\rho}_0^2 \alpha_k), \quad k = 1, 2, \ldots, \quad \alpha_1 = 2, \tag{6}$$

tandis que les paramètres de la méthode de stationnarisation à trois couches se définissent sur la base des formules

$$\tau_k \equiv \tilde{\tau}_0, \quad \alpha_k \equiv 1 + \tilde{\rho}_1^2, \quad k = 1, 2, \ldots, \quad \tilde{\rho}_1 = (1 - \sqrt{\tilde{\xi}})/(1 + \sqrt{\tilde{\xi}}). \tag{7}$$

Il s'ensuit de la théorie générale des méthodes itératives, exposée plus haut, que pour l'erreur $z_k = y_k - u$ des méthodes concernées sont vérifiées les estimations:

1) pour la méthode itérative simple

$$\| z_n \|_D \leqslant (\max_{\gamma_1 \leqslant t \leqslant \gamma_2} | 1 - \tilde{\tau}_0 t |)^n \, \| z_0 \|_D ; \tag{8}$$

2) pour la méthode à deux couches de Tchébychev et la méthode semi-itérative de Tchébychev

$$\| z_n \|_D \leqslant \tilde{q}_n \max_{\gamma_1 \leqslant t \leqslant \gamma_2} \left| T_n \left(\frac{1 - \tilde{\tau}_0 t}{\tilde{\rho}_0} \right) \right| \| z_0 \|_D, \tag{9}$$

où $\tilde{q}_n = 2\tilde{\rho}_1^n / (1 + \tilde{\rho}_1^{2n})$;

3) pour la méthode de stationnarisation à trois couches

$$\| z_n \|_D \leqslant \tilde{\rho}_1^n \max_{\gamma_1 \leqslant t \leqslant \gamma_2} \left| \frac{2\tilde{\rho}_1^2}{1 + \tilde{\rho}_1^2} T_n \left(\frac{1 - \tilde{\tau}_0 t}{\tilde{\rho}_0} \right) + \right.$$
$$\left. + \frac{1 - \tilde{\rho}_1^2}{1 + \tilde{\rho}_1^2} U_n \left(\frac{1 - \tilde{\tau}_0 t}{\tilde{\rho}_0} \right) \right| \| z_0 \|_D. \tag{10}$$

$T_n(x)$ et $U_n(x)$ sont ici des polynômes de Tchébychev de première et seconde espèces, γ_1 et γ_2 les valeurs précises des constantes dans (1).

Les estimations données définissent la vitesse de convergence des méthodes étudiées quand les paramètres d'itération sont calculés sur la base d'une information à priori imprécise.

2. Estimations de la vitesse de convergence des méthodes. Apprécions maintenant les maximums des modules des polynômes entrant dans les estimations (8)-(10). Opérons pour cela dans (8)-(10) à une substitution en posant $x = (1 - \tilde{\tau}_0 t)/\tilde{\rho}_0$ et posons $a = (1 - \tilde{\tau}_0\gamma_2)\tilde{\rho}_0$, $b = (1 - \tilde{\tau}_0\gamma_1)/\tilde{\rho}_0$. Alors les estimations (8)-(10) prendront la forme

$$\| z_n \|_D \leqslant \tilde{\rho}_0^n \left(\max_{a\leqslant x\leqslant b} | x |\right)^n \| z_0 \|_D,$$

$$\| z_n \|_D \leqslant \tilde{q}_n \max_{a\leqslant x\leqslant b} | T_n (x) | \, \| z_0 \|_D, \tag{11}$$

$$\| z_n \|_D \leqslant \tilde{\rho}_1^n \max_{a\leqslant x\leqslant b} \left| \frac{2\tilde{\rho}_1^2}{1+\tilde{\rho}_1^2} T_n (x) + \frac{1-\tilde{\rho}_1^2}{1+\tilde{\rho}_1^2} U_n (x) \right| \| z_0 \|_D.$$

Voyons d'abord le cas où $\tilde{\gamma}_1$ et $\tilde{\gamma}_2$ sont des approximations de γ_1 et γ_2 respectivement par le bas et par le haut, c'est-à-dire

$$\tilde{\gamma}_1 \leqslant \gamma_1 \leqslant \gamma_2 \leqslant \tilde{\gamma}_2. \tag{12}$$

Dans ce cas, comme il est aisé de le vérifier, les inégalités $-1 \leqslant \leqslant a \leqslant b \leqslant 1$ seront satisfaites. A partir de (11) il s'ensuit que la vitesse de convergence de la méthode itérative simple s'appréciera au moyen de la quantité $\tilde{\rho}_0^n$, de la méthode de Tchébychev à deux couches et de la méthode semi-itérative de Tchébychev au moyen de la quantité $\tilde{q}_n$, et de la méthode de stationnarisation à trois couches au moyen de la quantité $\tilde{\rho}_1^n (1 + n (1 - \tilde{\rho}_1^2)/(1 + \tilde{\rho}_1^2))$. Les méthodes itératives convergeront mais la vitesse de convergence décroîtra.

Examinons l'exemple pour lequel sont remplies les conditions (12). Soient

$$\tilde{\gamma}_1 = \gamma_1 (1 - \alpha), \quad \tilde{\gamma}_2 = \gamma_2, \quad 0 \leqslant \alpha < 1.$$

Dans ce cas $a = -1$, $b < 1$. Aussi pour l'erreur des méthodes étudiées on obtient de (11) les estimations suivantes:

$$\| z_n \|_D \leqslant \tilde{\rho}_0^n \| z_0 \|_D,$$

$$\| z_n \|_D \leqslant \tilde{q}_n \| z_0 \|_D,$$

$$\| z_n \|_D \leqslant \tilde{\rho}_1^n (1 + n (1 - \tilde{\rho}_1^2)/(1 + \tilde{\rho}_1^2)) \| z_0 \|_D.$$

A partir des formules correspondantes pour le nombre d'itérations on obtient que pour la méthode itérative simple, au cas d'une fixation imprécise de γ_1, ce nombre augmente d'environ $1/(1 - \alpha)$ fois par rapport à la fixation précise de γ_1, tandis que pour la méthode de Tchébychev à deux couches et la méthode semi-itérative de Tchébychev ce nombre ne s'accroît que de $1/\sqrt{1 - \alpha}$ fois.

Supposons maintenant que les conditions (12) ne sont pas remplies. Dans ce cas $\max (| a |, | b |) > 1$. Introduisons les notations suivantes

$$\frac{1}{\rho_0^*} = \max (| a |, | b |),$$

$$q_n^* = \frac{1}{T_n\left(\dfrac{1}{\rho_0^*}\right)} = \frac{2\rho_1^{*n}}{1+\rho_1^{*2n}}, \quad \rho_1^* = \frac{\rho_0^*}{1+\sqrt{1-\rho_0^{*2}}}.$$

En utilisant ces notations, de même que le rapport entre les polynômes de Tchébychev de première et seconde espèces

$$U_n(x) = (T^2_{n+1}(x) - 1)^{1/2}/(T^2_1(x) - 1)^{1/2}, \quad |x| \geqslant 1,$$

il vient

$$\max_{a \leqslant x \leqslant b} |x| = \frac{1}{\rho_0^*}, \quad \max_{a \leqslant x \leqslant b} |T_n(x)| = T_n\left(\frac{1}{\rho_0^*}\right) = \frac{1}{q_n^*} \leqslant \frac{1}{\rho_1^{*n}},$$

$$\max_{a \leqslant x \leqslant b} |U_n(x)| = U_n\left(\frac{1}{\rho_0^*}\right) = \frac{1 - \rho_1^{*2(n+1)}}{\rho_1^{*n}(1 - \rho_1^{*2})} \leqslant \frac{n+1}{\rho_1^{*n}}.$$

En portant ces estimations dans (11), on obtient

$$\| z_n \|_D \leqslant \left(\frac{\tilde{\rho}_0}{\rho_0^*}\right)^n \| z_0 \|_D, \tag{13}$$

$$\| z_n \|_D \leqslant \frac{\tilde{q}_n}{q_n^*} \| z_0 \|_D, \tag{14}$$

$$\| z_n \|_D \leqslant (\tilde{\rho}_1/\rho_1^*)^n (1 + n(1 - \tilde{\rho}_1^2)/(1 + \tilde{\rho}_1^2)) \| z_0 \|_D. \tag{15}$$

Notons que si H est un espace de dimension finie, on est en mesure d'indiquer l'approximation initiale y_0 pour laquelle dans les estimations (13), (14) on aboutira à des égalités.

Cherchons maintenant la condition avec la satisfaction de laquelle il est possible de garantir la convergence des méthodes itératives étudiées construites sur la base d'une information à priori imprécise. Comme le rapport $\tilde{q}_n/q_n^*$ ne tend vers zéro pour $n \to \infty$ qu'à la condition que $\rho_1^* > \tilde{\rho}_1$, cette condition étant équivalente à l'exigence de $\rho_0^* > \tilde{\rho}_0$, il s'ensuit donc de (13)-(15) que les méthodes itératives convergeront si est satisfaite l'inégalité

$$\tilde{\rho}_0 < \rho_0^*. \tag{16}$$

En utilisant les définitions de ρ_0^*, a et b, on constate que (16) se vérifie pour $|1 - \tilde{\tau}_0\gamma_1| < 1$, $|1 - \tilde{\tau}_0\gamma_2| < 1$. En résolvant ces inégalités, on obtient

$$\tilde{\gamma}_1 + \tilde{\gamma}_2 > \gamma_2. \tag{17}$$

Bref, si la condition (17) est satisfaite, les méthodes itératives construites sur la base d'une information à priori imprécise convergeront. Il s'ensuit de ce qui vient d'être dit qu'au cas d'un espace H de dimension finie la condition (17) est aussi une condition nécessaire de la convergence des méthodes.

Apprécions maintenant le nombre réel d'itérations nécessaires à l'obtention de la précision exigée ε. Désignons, comme auparavant, par n le nombre d'itérations au cas de la fixation précise de l'information à priori et par $\tilde{n}$ le nombre théorique d'itérations calculé au moyen des formules des théorèmes correspondants se rapportant à une information à priori imprécise, n^* désignant le nombre d'itérations réel qui permet d'atteindre la précision ε. Il s'ensuit des formules (13)-(15) que le nombre d'itérations réel n^* doit découler des conditions :

1) pour la méthode itérative simple, de la condition $\tilde{\rho}_0^n \leqslant \varepsilon\rho_0^{*n}$;

2) pour la méthode de Tchébychev à deux couches et la méthode semi-itérative de Tchébychev, de la condition $\tilde{q}_n \leqslant \varepsilon q_n^*$.

On se convainc sans peine qu'on a les inégalités $n^* \geqslant \tilde{n}$, $n^* \geqslant n$, le nombre d'itérations $\tilde{n}$ pouvant être plus grand ou plus petit que n. Etant donné que la seule caractéristique quantitative de la méthode itérative pouvant être calculée est le nombre théorique d'itérations $\tilde{n}$, il est important pour la mise en œuvre de la méthode d'apprécier de combien de fois le nombre réel n^* sera plus grand que $\tilde{n}$. Pour la comparaison théorique de la qualité des méthodes itératives, il faut pouvoir apprécier le rapport n^*/n.

Cherchons les estimations exigées pour un exemple. Soient $\tilde{\gamma}_1$ et $\tilde{\gamma}_2$ des approximations pour γ_1 et γ_2 respectivement par le haut et par le bas

$$\tilde{\gamma}_1 = (1+\alpha)\gamma_1, \quad \tilde{\gamma}_2 = (1-\alpha)\gamma_2, \quad \alpha \geqslant 0. \tag{18}$$

De la condition (17) et de l'exigence naturelle que $\tilde{\gamma}_1 \leqslant \tilde{\gamma}_2$ il vient que les méthodes convergeront si est satisfaite la condition

$$\alpha < \min\,(\xi/(1-\xi),\ (1-\xi)/(1+\xi)), \quad \xi = \gamma_1/\gamma_2.$$

Pour l'exemple pris auront lieu les inégalités $n^* \geqslant n \geqslant \tilde{n}$. En effet, de (18) on tire

$$\tilde{\xi} = \frac{\tilde{\gamma}_1}{\tilde{\gamma}_2} = \frac{1+\alpha}{1-\alpha}\xi \geqslant \xi$$

et, partant,

$$\tilde{\rho}_0 \leqslant \rho_0 = (1-\xi)/(1+\xi), \quad \tilde{\rho}_1 \leqslant \rho_1 = (1-\sqrt{\xi})/(1+\sqrt{\xi}), \quad \tilde{q}_n \leqslant q_n.$$

Il s'ensuit de là que $n \geqslant \tilde{n}$. Apprécions maintenant les grandeurs entrant dans les inégalités (13)-(14). Vu que

$$\tilde{\tau} = 2/(\tilde{\gamma}_1 + \tilde{\gamma}_2) = \tau_0/(1-\alpha\rho_0) < \tau_0, \quad \tau_0 = 2/(\gamma_1+\gamma_2),$$

on doit avoir

$$1/\rho_0^* = \max\,(|a|,\ |b|) = |a| = (\tilde{\tau}_0\gamma_2 - 1)/\tilde{\rho}_0.$$

En omettant les calculs peu compliqués, on obtient

$$\tilde{\rho}_0 = \frac{1-\tilde{\xi}}{1+\tilde{\xi}} = \frac{\rho_0 - \alpha}{1-\alpha\rho_0},$$

$$\frac{\tilde{\rho}_0}{\rho_0^*} = \tilde{\tau}_0\gamma_2 - 1 = 1 - \frac{1 - \frac{\alpha}{\xi}(1-\xi)}{1+\alpha}(1-\tilde{\rho}_0) = 1 - \frac{1-\frac{\alpha}{\xi}(1-\xi)}{1-\alpha\rho_0}(1-\rho_0),$$

$$\tilde{\rho}_1 = \frac{1-\sqrt{\tilde{\xi}}}{1+\sqrt{\tilde{\xi}}} = \frac{\rho_0 - \alpha}{1-\alpha\rho_0 + \sqrt{(1-\alpha^2)(1+\rho_0^2)}},$$

$$\frac{\tilde{\rho}_1}{\rho_1^*} = 1 - \frac{(1+\alpha)\sqrt{\xi} + \sqrt{1-\alpha^2} - \frac{\alpha}{\sqrt{\xi}} - \sqrt{\frac{\alpha}{\xi}[1-(1+\alpha)\xi]}}{(1+\alpha)\sqrt{\xi} + \sqrt{1-\alpha^2}}(1-\tilde{\rho}_1) =$$

$$= 1 - \frac{\left[(1+\alpha)\sqrt{\xi} + \sqrt{1-\alpha^2} - \frac{\alpha}{\sqrt{\xi}} - \sqrt{\frac{\alpha}{\xi}[1-(1+\alpha)\xi]}\right](1+\sqrt{\xi})}{1-\alpha+(1+\alpha)\xi + 2\sqrt{(1-\alpha^2)\xi}}(1-\rho_1).$$

Voyons d'abord la méthode itérative simple. Sur la base du théorème 2, § 3, ch. VI, et à partir de (13) on déduit pour les nombres d'itérations n^*, $\tilde{n}$ et n de la méthode itérative simple les estimations suivantes:

$$n=\frac{\ln\varepsilon}{\ln\rho_0}\approx\frac{\ln(1/\varepsilon)}{1-\rho_0},\quad \tilde{n}=\frac{\ln\varepsilon}{\ln\tilde{\rho}_0}\approx\frac{\ln(1/\varepsilon)}{1-\tilde{\rho}_0},$$

$$n^*=\frac{\ln\varepsilon}{\ln(\tilde{\rho}_0/\rho_0^*)}=\frac{\ln(1/\varepsilon)}{1-\tilde{\rho}_0/\rho_0^*}.$$

En y portant les expressions obtenues plus haut, on obtient

$$\frac{n^*}{\tilde{n}}\approx\frac{1+\alpha}{1-\dfrac{\alpha}{\xi}(1-\xi)},\quad \frac{n^*}{n}\approx\frac{1-\alpha\rho_0}{1-\dfrac{\alpha}{\xi}(1-\xi)}.$$

Si $\alpha\approx c\xi$, où $c<1$, on en tire

$$n^*\approx\tilde{n}/(1-c),\quad n^*\approx n/(1-c).$$

Bref, si $\alpha\approx c\xi$, le nombre réel d'itérations n^* pour la méthode itérative simple est $1/(1-c)$ fois plus grand que le nombre théorique d'itérations $\tilde{n}$ calculé sur la base d'une information à priori imprécise.

Passons maintenant à la méthode de Tchébychev et à la méthode semi-itérative de Tchébychev. Sur la base de la définition de $\tilde{q}_n$ et q_n^*, on obtient

$$\frac{\tilde{q}_n}{q_n^*}=\frac{\tilde{\rho}_1^n}{\rho_1^{*n}}\cdot\frac{1+\rho_1^{*2n}}{1+\tilde{\rho}_1^{2n}}\leqslant\frac{2(\tilde{\rho}_1/\rho_1^*)^n}{1+(\tilde{\rho}_1/\rho_1^*)^{2n}}.$$

On trouve donc pour le nombre d'itérations n^* l'estimation suivante:

$$n^*=\frac{\ln(0{,}5\varepsilon)}{\ln(\tilde{\rho}_1/\rho_1^*)}\approx\frac{\ln(2/\varepsilon)}{1-\tilde{\rho}_1/\rho_1^*}.$$

Ensuite, sur la base du théorème 1, § 2, ch. VI, et du théorèm 1, § 2, ch. VII, on déduit les estimations pour n et $\tilde{n}$:

$$n=\frac{\ln(0{,}5\varepsilon)}{\ln\rho_1}\approx\frac{\ln(2/\varepsilon)}{1-\rho_1},\quad \tilde{n}=\frac{\ln(0{,}5\varepsilon)}{\ln\tilde{\rho}_1}\approx\frac{\ln(2/\varepsilon)}{1-\tilde{\rho}_1}.$$

En y portant les expressions obtenues plus haut pour le rapport $\tilde{\rho}_1/\rho_1^*$ et en posant que $\alpha\approx c\xi$, on trouve

$$n^*/\tilde{n}\approx 1/(1-\sqrt{c}),\quad n^*/n\approx 1/(1-\sqrt{c}).$$

Donc si $\alpha\approx c\xi$, où $c<1$, le nombre réel d'itérations n^* pour la méthode de Tchébychev et la méthode semi-itérative de Tchébychev est environ $1/(1-\sqrt{c})$ fois plus grand que le nombre théorique d'itérations $\tilde{n}$ calculé sur la base d'une information à priori imprécise.

CHAPITRE VIII

MÉTHODES ITÉRATIVES DU TYPE VARIATIONNEL

On étudie dans ce chapitre les méthodes itératives à deux et trois couches du type variationnel. Pour la mise en œuvre de ces méthodes on peut se dispenser de toute information à priori sur les opérateurs du schéma itératif. Dans les §§ 1, 2 on étudie les méthodes du gradient à deux couches et dans les §§ 3, 4 les méthodes à trois couches de directions conjuguées. L'accélération de la convergence des méthodes à deux couches au cas d'autoconjugaison est traitée au § 5

§ 1. Méthode du gradient à deux couches

1. Position du problème sur le choix des paramètres d'itération. Pour trouver la solution approchée de l'équation linéaire opératorielle

$$Au = f \tag{1}$$

avec opérateur A non dégénéré et associé à l'espace hilbertien réel H, examinons le schéma itératif implicite à deux couches

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \ldots, \tag{2}$$

avec l'approximation initiale arbitraire $y_0 \in H$ et l'opérateur B non dégénéré.

Le schéma itératif (2) a déjà été étudié au chapitre VI, où on a construit les jeux des paramètres d'itération $\{\tau_k\}$ et fourni les estimations de la vitesse de convergence des méthodes itératives correspondantes (de la méthode de Tchébychev et de la méthode itérative simple).

Toute méthode itérative à deux couches, construite sur la base du schéma (2), peut être caractérisée par les opérateurs A et B, l'espace énergétique H_D, dans lequel on démontre la convergence de la méthode, et le jeu des paramètres d'itération τ_k. Le problème principal de la théorie des méthodes itératives réside dans le choix optimal des paramètres τ_k.

Dans le chapitre VI les méthodes itératives ont été construites avec les paramètres τ_k choisis sur la base de la condition du minimum dans H_D soit de la norme de l'opérateur de transfert d'une itération à l'autre, soit de la norme de l'opérateur résolvant. Le

trait distinctif des méthodes itératives construites sur la base de ce principe consiste dans l'utilisation pour le calcul des paramètres τ_k d'une information à priori déterminée sur les opérateurs du schéma itératif.

L'aspect de l'information à priori est fonction des propriétés des opérateurs A, B et D. Au cas où l'opérateur $DB^{-1}A$ est autoadjoint dans l'espace H, cette information se réduit à la fixation des constantes de l'équivalence énergétique des opérateurs D et $DB^{-1}A$, autrement dit des constantes γ_1 et γ_2 des inégalités

$$\gamma_1 D \leqslant DB^{-1}A \leqslant \gamma_2 D, \quad \gamma_1 > 0, \tag{3}$$

ou des bornes de l'opérateur $DB^{-1}A$ dans H_D.

Au cas de non-autoconjugaison on utilise soit deux nombres γ_1 et γ_2 des inégalités

$$\gamma_1 D \leqslant DB^{-1}A, \quad (DB^{-1}Ax, \ B^{-1}Ax) \leqslant \gamma_2 (DB^{-1}Ax, x), \quad \gamma_1 > 0, \tag{4}$$

soit trois nombres γ_1, γ_2 et γ_3, où γ_1 et γ_2 sont des constantes des inégalités (3) et γ_3 une constante de l'inégalité

$$\| 0{,}5 \, (DB^{-1}A - A^* (B^*)^{-1} D) \, x \|^2_{D^{-1}} \leqslant \gamma_3^2 \, (Dx, x), \tag{5}$$

ou de l'inégalité

$$\| 0{,}5 \, (DB^{-1}A - A^* (B^*)^{-1} D) \, x \|^2_{D^{-1}} \leqslant \gamma_3 \, (DB^{-1}Ax, \ x). \tag{6}$$

En maintes occasions la recherche des constantes γ_1, γ_2 et γ_3 avec une suffisante précision peut s'avérer compliquée et constituer un problème séparé dont la résolution exigera le recours à des méthodes de calcul spéciales. Si l'information à priori peut être obtenue par des calculs peu laborieux ou s'il faut résoudre une série de problèmes (1) aux seconds membres différents, il est rationnel de rechercher une fois pour toutes l'information à priori nécessaire et, ensuite, recourir aux méthodes itératives construites au chapitre VI. Ce procédé est recommandé si le temps complémentaire dépensé à l'obtention de l'information à priori est de beaucoup inférieur à celui exigé pour la résolution de toute la série de problèmes (1).

Au cas où il s'agit de ne résoudre qu'un problème (1) ou quand l'approximation initiale est donnée de façon très correcte, tandis que le calcul des constantes γ_1, γ_2 et γ_3 est un processus fort laborieux, il faut recourir aux méthodes itératives du type variationnel dont on va aborder l'étude.

Dans les méthodes itératives à deux couches du type variationnel on n'a pas besoin, pour le calcul des paramètres τ_k, de recourir à une information à priori quelconque sur les opérateurs du schéma (2) (à part les conditions de forme générale $A = A^* > 0$, $(DB^{-1}A)^* = DB^{-1}A$, etc.), la construction de ces méthodes s'appuyant sur le principe suivant. Si l'approximation y_k est donnée, tandis que y_{k+1} s'obtient suivant le schéma (2), le paramètre d'itération τ_{k+1} est alors choisi sur la base de la condition du minimum

dans H_D de la norme d'erreur $z_{k+1} = y_{k+1} - u$, où u est la solution de l'équation (1).

La dénomination des méthodes est liée au fait que la suite y_k construite suivant la formule (2) et, où les paramètres τ_k sont choisis sur la base de la condition mentionnée plus haut, est une suite minimisante pour la fonctionnelle quadratique

$$I(y) = (D(y-u),\ y-u).$$

Cette fonctionnelle, en vertu de la définissabilité positive de l'opérateur D, est bornée par le bas et atteint un minimum égal à zéro sur la solution de l'équation (1), c'est-à-dire pour $y = u$. Le choix du paramètre τ_{k+1} sur la base de la condition mentionnée garantit la minimisation locale de la fonctionnelle $I(y)$ avec le passage de y_k à y_{k+1}, c'est-à-dire en un pas itératif. Au cas d'un schéma explicite ($B = E$) le passage de y_k à y_{k+1} est réalisé suivant la formule

$$y_{k+1} = y_k - \tau_{k+1} r_k, \quad r_k = Ay_k - f.$$

Notons que pour un opérateur autoadjoint défini positif A le passage de y_k à y_{k+1} s'effectue suivant la direction $-r_k$ qui coïncide avec celle de l'antigradient de la fonctionnelle $(A(y-u),\ y-u)$ au point y_k. On sait que le décroissement maximal de la fonctionnelle s'effectue suivant la direction de l'antigradient. Aussi ces méthodes sont-elles quelquefois appelées méthodes de descente par gradient ou tout simplement méthodes du gradient. On conservera également cette dénomination pour les méthodes implicites à deux couches du type variationnel.

Notre premier objectif est de trouver le paramètre τ_{k+1} sur la base de la condition du minimum dans H_D de la norme d'erreur $z_{k+1} = y_{k+1} - u$.

2. Formule pour paramètres d'itération. Cherchons maintenant la formule pour le calcul du paramètre d'itération τ_{k+1} en posant que l'opérateur A n'est pas dégénéré. Ecrivons d'abord l'équation de l'erreur $z_k = y_k - u$, $k = 0, 1, \ldots$ En portant $y_k = z_k + u$ dans le schéma (2), on obtient

$$z_{k+1} = (E - \tau_{k+1} B^{-1} A)\, z_k, \quad k = 0, 1, \ldots, \quad z_0 = y_0 - u.$$

La substitution $z_k = D^{-1/2} x_k$ permet de passer à l'équation ne comprenant qu'un seul opérateur

$$x_{k+1} = S_{k+1} x_k, \quad S_k = E - \tau_k C,$$
$$C = D^{-1/2} (DB^{-1}A)\, D^{-1/2}. \tag{7}$$

En utilisant l'égalité $\| z_k \|_D = \| x_k \|$, on peut formuler le problème posé plus haut du choix du paramètre τ_{k+1} de la façon suivante: choisirons le paramètre τ_{k+1} sur la base de la condition du minimum de la norme x_{k+1} dans l'espace H.

Résolvons ce problème. Calculons la norme x_{k+1}:

$$\|x_{k+1}\|^2 = ((E-\tau_{k+1}C)\,x_k, \quad (E-\tau_{k+1}C)\,x_k) =$$
$$= \|x_k\|^2 - 2\tau_{k+1}(Cx_k, x_k) + \tau_{k+1}^2 (Cx_k, Cx_k) =$$
$$= (Cx_k, Cx_k)\left[\tau_{k+1} - \frac{(Cx_k, x_k)}{(Cx_k, Cx_k)}\right]^2 + \|x_k\|^2 - \frac{(Cx_k, x_k)^2}{(Cx_k, Cx_k)}.$$

L'opérateur A étant non dégénéré, l'opérateur C ne l'est également pas. Aussi pour tout x_k a-t-on $(Cx_k, Cx_k) > 0$, et le minimum de la norme x_{k+1} est atteint pour

$$\tau_{k+1} = \frac{(Cx_k, x_k)}{(Cx_k, Cx_k)}. \tag{9}$$

En portant (9) dans (8), il vient

$$\|x_{k+1}\| = \rho_{k+1}\|x_k\|, \tag{10}$$

où

$$\rho_{k+1}^2 = 1 - \frac{(Cx_k, x_k)^2}{(Cx_k, Cx_k)(x_k, x_k)}. \tag{11}$$

Bref, la formule (9) définit la valeur optimale du paramètre d'itération τ_{k+1}. En portant dans (9) $x_k = D^{1/2}z_k$, il vient

$$\tau_{k+1} = \frac{(DB^{-1}Az_k, z_k)}{(DB^{-1}Az_k, B^{-1}Az_k)}, \quad k = 0, 1, \ldots$$

Compte tenu de ce que $Az_k = Ay_k - Au = Ay_k - f = r_k$ est le résidu, tandis que $B^{-1}r_k = w_k$ est la correction, la formule pour le paramètre τ_{k+1} peut être écrite sous la forme suivante:

$$\tau_{k+1} = \frac{(Dw_k, z_k)}{(Dw_k, w_k)}, \quad k = 0, 1, \ldots, \tag{12}$$

tandis que le schéma itératif (2) s'écrit sous forme de formule explicite pour le calcul de y_{k+1}:

$$y_{k+1} = y_k - \tau_{k+1}w_k, \quad k = 0, 1, \ldots \tag{13}$$

L'algorithme mettant en œuvre la méthode construite peut être décrit de la façon suivante:

1) sur la base de y_k donné on calcule le résidu $r_k = Ay_k - f$,
2) on résout l'équation de la correction $Bw_k = r_k$,
3) suivant la formule (12) on calcule le paramètre τ_{k+1},
4) suivant la formule (13) on obtient la nouvelle approximation y_{k+1}.

Les formules (12) ne peuvent encore servir aux calculs car à côté des quantités r_k et w_k, connues du fait du processus d'itération, elles contiennent l'erreur inconnue z_k. Au § 2, en choisissant un opérateur D concret, on obtiendra des formules pour les paramètres τ_k où ne figureront que des quantités connues. Passons, en attendant, à l'appréciation de l'estimation de la vitesse de convergence de la méthode itérative construite.

3. Appréciation de la vitesse de convergence. Apprécions maintenant la vitesse de convergence des méthodes du gradient à deux couches. Etant donné que le paramètre d'itération τ_{k+1} est choisi sur la base de la condition du minimum dans H_D de la norme d'erreur z_{k+1}, équivalente à la condition du minimum dans H de la norme x_{k+1}, il s'ensuit de (7) que

$$\| x_{k+1} \| = \min_{\tau_{k+1}} \| S_{k+1} x_k \| \leqslant \min_{\tau_{k+1}} \| S_{k+1} \| \| x_k \| =$$
$$= \min_{\tau} \| E - \tau C \| \| x_k \| = \rho \| x_k \|, \quad \rho = \min_{\tau} \| E - \tau C \|.$$

En comparant cette estimation à l'inégalité (10), on obtient

$$\rho_k \leqslant \rho \leqslant 1, \quad k = 1, 2, \ldots \tag{14}$$

A partir de (10), (14) on tire l'estimation $\| x_{k+1} \| \leqslant \rho \| x_k \|$, et en vertu de la substitution effectuée $x_k = D^{1/2} z_k$ il s'ensuit l'estimation de la norme d'erreur z_n dans l'espace énergétique H_D:

$$\| z_n \|_D \leqslant \rho^n \| z_0 \|_D, \quad \rho = \min_{\tau} \| E - \tau C \|. \tag{15}$$

Si la condition $\rho < 1$ est remplie, la méthode du gradient à deux couches converge dans H_D. Il s'ensuit de l'estimation (15) que pour diminuer la norme de l'erreur initiale dans H_D de $1/\varepsilon$ fois il suffit d'effectuer $n \geqslant n_0(\varepsilon)$ itérations, où

$$n_0(\varepsilon) = \ln \varepsilon / \ln \rho. \tag{16}$$

Ainsi la vitesse de convergence de la méthode du gradient à deux couches se définit par la quantité ρ. Rappelons que dans le chapitre VI, en étudiant la méthode itérative simple avec des hypothèses variées sur l'opérateur C, on a obtenu des estimations de ρ. La valeur ρ définit la vitesse de convergence de la méthode itérative simple. Donc de l'estimation (15) obtenue ici il s'ensuit que toute méthode du gradient à deux couches converge à une vitesse non moindre que la méthode itérative simple.

Donnons les estimations pour ρ obtenues aux §§ 3, 4, ch. VI pour des hypothèses variées sur les opérateurs A, B et D.

1. Si l'opérateur $DB^{-1}A$ est autoadjoint dans H et γ_1 et γ_2 sont les constantes des inégalités (3), on a alors

$$\rho = (1 - \xi)/(1 + \xi), \quad \xi = \gamma_1/\gamma_2. \tag{17}$$

2. Soit un opérateur $DB^{-1}A$ non autoadjoint dans H;

a) si la condition (4) est remplie, on a

$$\rho = \sqrt{1 - \xi}, \quad \xi = \gamma_1/\gamma_2; \tag{18}$$

b) si c'est les conditions (3), (5) qui sont remplies, on a

$$\rho = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{1-\varkappa}{1+\varkappa}\frac{\gamma_1}{\gamma_2}, \quad \varkappa = \frac{\gamma_3}{\sqrt{\gamma_1\gamma_2 + \gamma_3^2}}. \tag{19}$$

On a ainsi démontré le

Théorème 1. *Si au cas du schéma* (2) *la méthode itérative simple converge, il y a également convergence de la méthode du gradient à deux couches* (2), (12). *De plus, pour l'erreur z_n se vérifie l'estimation*

$$\| z_n \|_D \leqslant \rho^n \| z_0 \|_D,$$

où ρ est défini dans (17), *si l'opérateur $DB^{-1}A$ est autoadjoint dans H et les conditions* (3) *sont remplies, ρ est défini dans* (18) *si pour l'opérateur non autoadjoint $DB^{-1}A$ les conditions* (4) *sont remplies et dans* (19) *si les conditions* (3), (15) *sont remplies. L'estimation du nombre d'itérations est donnée dans* (16).

Remarque. Si l'équation (1) est étudiée dans l'espace hilbertien complexe, on doit alors choisir le paramètre d'itération τ_{k+1} suivant la formule

$$\tau_{k+1} = \frac{\operatorname{Re}(Dw_k, z_k)}{(Dw_k, w_k)}, \quad k = 0, 1, \ldots$$

Le théorème 1 reste valable, mais les conditions (3), (4) doivent être remplacées par les inégalités

$$\gamma_1 (Dx, x) \leqslant \operatorname{Re} (DB^{-1}Ax, x) \leqslant \gamma_2 (Dx, x),$$
$$\gamma_1 (Dx, x) \leqslant \operatorname{Re} (DB^{-1}Ax, x),$$
$$(DB^{-1}Ax, B^{-1}Ax) \leqslant \gamma_2 \operatorname{Re} (DB^{-1}Ax, x),$$

où Re z est la partie réelle du nombre complexe z.

4. Impossibilité d'améliorer l'estimation au cas d'opérateurs autoadjoints. Montrons que pour la classe d'approximations initiales y_0 quelconques au cas d'un opérateur $DB^{-1}A$ autoadjoint dans l'espace de dimension finie H l'estimation à priori de l'erreur de la méthode itérative (2), (12), obtenue au théorème 1, ne peut être améliorée. Pour cela il suffit d'indiquer une telle approximation initiale x_0 pour laquelle la résolution de l'équation (7) implique l'égalité $\| x_{k+1} \| = \rho \|x_k\|$, où ρ est défini dans (17).

Cherchons l'approximation initiale x_0. Soit H l'espace de dimension finie ($H = H_N$). Vu que l'opérateur $DB^{-1}A$ est autoadjoint dans H, l'opérateur $C = D^{-1/2} (DB^{-1}A) D^{-1/2}$ l'est également dans H. Il existe donc un système complet des fonctions propres $v_1, v_2, \ldots, v_N$ de l'opérateur C. Désignons par λ_k la valeur propre de l'opérateur C correspondant à la fonction propre v_k, de sorte que $Cv_k = \lambda_k v_k$, $k = 1, 2, \ldots, N$. Posons $\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_N$. Comme les inégalités (3) sont équivalentes aux inégalités

$$\gamma_1 E \leqslant C \leqslant \gamma_2 E, \quad \gamma_1 > 0,$$

dans (3) au lieu de γ_1 et γ_2 on peut prendre λ_1 et λ_N. En outre, ρ, défini dans (17), peut être écrit sous la forme : $\rho = (\lambda_N - \lambda_1)/(\lambda_N + \lambda_1)$. Choisissons l'approximation initiale

$$x_0 = \sqrt{\lambda_N}\, v_1 + \sqrt{\lambda_1}\, v_N. \tag{20}$$

On a alors $Cx_0 = \lambda_1 \sqrt{\lambda_N} v_1 + \lambda_N \sqrt{\lambda_1} v_N$. Profitant de l'orthonormalisation du système des fonctions propres $v_1, v_2, \ldots, v_N$, il vient

$$(x_0, x_0) = \lambda_1 + \lambda_N,$$
$$(Cx_0, x_0) = 2\lambda_1\lambda_N,$$
$$(Cx_0, Cx_0) = \lambda_1\lambda_N (\lambda_1 + \lambda_N).$$

En portant ces valeurs dans (9), (11), on obtient $\tau_1 = 2/(\lambda_1 + \lambda_N)$, $\rho_1 = (\lambda_N - \lambda_1)/(\lambda_N + \lambda_1) = \rho$. Il s'ensuit de (10) l'égalité $\| x_1 \| = \rho \| x_0 \|$, tandis qu'à partir de (7) on obtient x_1:

$$x_1 = \rho \left(\sqrt{\lambda_N}\, v_1 - \sqrt{\lambda_1}\, v_N \right).$$

En poursuivant les calculs, on obtient

$$Cx_1 = \rho \left(\lambda_1 \sqrt{\lambda_N}\, v_1 - \lambda_N \sqrt{\lambda_1}\, v_N \right),$$
$$(x_1, x_2) = \rho^2 (x_0, x_0),$$
$$(Cx_1, x_1) = \rho^2 (Cx_0, x_0),$$
$$(Cx_1, Cx_1) = \rho^2 (Cx_0, Cx_0).$$

Par conséquent,

$$\tau_2 = \frac{(Cx_1, x_1)}{(Cx_1, Cx_1)} = \frac{(Cx_0, x_0)}{(Cx_0, Cx_0)} = \tau_1,$$

$$\rho_2^2 = 1 - \frac{(Cx_1, x_1)^2}{(Cx_1, Cx_1)(x_1, x_1)} = 1 - \frac{(Cx_0, x_0)^2}{(Cx_0, Cx_0)(x_0, x_0)} = \rho_1^2 = \rho^2.$$

Donc $\| x_2 \| = \rho \| x_1 \|$. En outre, $x_2 = x_1 - \tau_2 C x_1 = \rho^2 x_0$, c'est-à-dire x_2 est proportionnel à x_0. Il s'ensuit aussitôt que $\tau_3 = \tau_2 = \tau_1$, $\rho_3 = \rho$ et $x_3 = \rho^2 x_1$. Aussi pour tout k a-t-on:

$$\tau_k \equiv 2/(\lambda_1 + \lambda_N), \quad \rho_k \equiv \rho = (\lambda_N - \lambda_1)/(\lambda_1 + \lambda_1),$$
$$\| x_{k+1} \| = \rho \| x_k \|.$$

La proposition est démontrée.

On a ainsi montré que si l'approximation initiale est choisie suivant la formule (20), dans la méthode du gradient à deux couches tous les paramètres τ_k sont alors identiques et coïncident avec le paramètre de la méthode itérative simple (voir § 3, ch. VI), les erreurs sont proportionnelles toutes les deux itérations, tandis que la vitesse de convergence de la méthode est la plus lente.

Notons qu'une telle lenteur de la convergence n'a lieu pour la méthode qu'au cas d'une approximation initiale particulièrement « mauvaise ». Pour une « bonne » approximation initiale la vitesse de convergence de la méthode peut augmenter sensiblement. L'étude plus détaillée de la nature des variations de la vitesse de convergence

de la méthode fera l'objet du point suivant, en attendant on fournira un exemple illustrant la remarque donnée plus haut.

Montrons que si en qualité d'approximation initiale x_0 on choisit une fonction propre quelconque v_m, la méthode du gradient à deux couches convergera alors au bout d'une seule itération.

En effet, soit $x_0 = v_m$. Alors des calculs peu laborieux donnent

$$Cx_0 = \lambda_m v_m = \lambda_m x_0, \quad (Cx_0, x_0) = \lambda_m (x_0, x_0),$$
$$(Cx_0, Cx_0) = \lambda_m^2 (x_0, x_0), \quad \tau_1 = 1/\lambda_m, \quad \rho_1 = 0,$$

c'est-à-dire $x_1 = 0$ ou $y_1 = u$.

Cette propriété qualitativement nouvelle des méthodes du gradient à deux couches, qui leur permet d'accroître la vitesse de convergence au cas de choix d'une « bonne » approximation initiale, distingue ces méthodes des méthodes itératives à deux couches étudiées au chapitre VI et orientées de façon stricte sur le choix de la plus mauvaise approximation initiale.

5. Propriété asymptotique des méthodes du gradient au cas d'opérateurs autoadjoints. Passons maintenant à la propriété asymptotique des méthodes du gradient à deux couches que ces dernières possèdent quand l'opérateur $DB^{-1}A$ est autoadjoint. Cette propriété réside dans le fait que la suite $\{\rho_k\}$, définie dans (11), est croissante. Etant donné que la quantité ρ_k détermine la vitesse de décroissance de la norme d'erreur avec le passage de la k-ième à la $(k+1)$-ième itération, la présence de cette propriété implique une diminution de la norme d'erreurs z_n pour des n grands par rapport au début du processus d'itérations. De plus, pour des n suffisamment grands la convergence des méthodes du gradient devient pratiquement identique à celle de la méthode itérative simple.

On montrera que pour des grands numéros d'itérations les erreurs deviennent toutes les deux itérations presque proportionnelles. En utilisant ce fait, on construira une méthode approchée d'obtention des constantes γ_1 et γ_2 des inégalités (3) et, au § 5, on construira le processus d'accélération de la convergence des méthodes du gradient à deux couches.

Admettons donc que l'opérateur $DB^{-1}A$ et avec lui l'opérateur C sont autoadjoints dans H. Montrons que la suite $\{\rho_k\}$ est croissante. De (10) on déduit les égalités

$$\| x_{k+2} \| = \rho_{k+2} \| x_{k+1} \|, \quad \| x_{k+1} \| = \rho_{k+1} \| x_k \|.$$

Calculons la norme de la différence $x_{k+2} - \rho_{k+2}\rho_{k+1}x_k$:

$$\| x_{k+2} - \rho_{k+2}\rho_{k+1}x_k \|^2 = \| x_{k+2} \|^2 - 2\rho_{k+2}\rho_{k+1} (x_{k+2}, x_k) +$$
$$+ \rho_{k+2}^2\rho_{k+1}^2 \| x_k \|^2 = 2 (\| x_{k+2} \|^2 - \rho_{k+2}\rho_{k+1} (x_{k+2}, x_k)). \quad (21)$$

Calculons séparément le produit scalaire (x_{k+2}, x_k). De (7) on tire

$$x_{k+2} = x_{k+1} - \tau_{k+2}Cx_{k+1}, \quad x_k = x_{k+1} + \tau_{k+1}Cx_k. \quad (22)$$

En multipliant scalairement cette dernière égalité par Cx_k et compte tenu de (9), il vient

$$(Cx_k, x_k) = (x_{k+1}, Cx_k) + \tau_{k+1}(Cx_k, Cx_k) = \\ = (x_{k+1}, Cx_k) + (Cx_k, x_k).$$

Par conséquent, pour tout k on a l'égalité

$$(x_{k+1}, Cx_k) = 0, \tag{23}$$

tandis qu'en vertu de l'opérateur C qui est autoadjoint on aboutit à l'égalité $(Cx_{k+1}, x_k) = 0$.

De (22) et (23) on obtient

$$(x_{k+2}, x_k) = (x_{k+1} - \tau_{k+2}Cx_{k+1}, x_k) = (x_{k+1}, x_k) = \\ = (x_{k+1}, x_{k+1} + \tau_{k+1}Cx_k) = \| x_{k+1} \|^2.$$

En portant l'égalité obtenue dans (21), il vient

$$\| x_{k+2} - \rho_{k+2}\rho_{k+1}x_k \|^2 = 2\left(1 - \frac{\rho_{k+1}}{\rho_{k+2}}\right) \| x_{k+2} \|^2. \tag{24}$$

Il s'ensuit de (24) que soit $\rho_{k+2} > \rho_{k+1}$, soit $\rho_{k+1} = \rho_{k+2} = \bar{\rho}$ et $x_{k+2} = \bar{\rho}^2 x_k$. Dans ce dernier cas il est évident que pour tous $n \geqslant k$ on aura les égalités

$$\rho_{n+1} = \bar{\rho}, \quad x_{n+2} = \bar{\rho}^2 x_n, \tag{25}$$

autrement dit, la suite ρ_k tend vers la valeur limite.

Bref, on a montré que la suite $\{\rho_k\}$ est en fait croissante. Au point 3 de ce paragraphe on a montré que cette suite est bornée par le haut et, partant, possède une limite. Aussi pour des numéros k suffisamment grands aura-t-on l'égalité approchée $\rho_{k+1} \approx \rho_{k+2}$ et, par suite, $x_{k+2} \approx \rho_{k+2}\rho_{k+1}x_k$, autrement dit, toutes les deux itérations, les erreurs seront presque proportionnelles.

Examinons ce qu'il s'ensuivra de la tendance de la suite ρ_k vers une valeur limite. Dans ce cas on a les égalités (25), c'est-à-dire $x_{n+2} = \bar{\rho}^2 x_n$. Supposons que l'espace H est de dimension finie, $v_1, v_2, \ldots, v_N$ étant un système des fonctions propres de l'opérateur C. Développons x_n en fonctions propres

$$x_n = \sum_{k=1}^{N} \alpha_k^{(n)} v_k. \tag{26}$$

De l'équation (7) il vient

$$x_{n+2} = (E - \tau_{n+2}C)(E - \tau_{n+1}C)x_n = \\ = \sum_{k=1}^{N} (1 - \tau_{n+2}\lambda_k)(1 - \tau_{n+1}\lambda_k)\lambda_k^{(n)} v_k.$$

Comme $x_{n+2} = \bar{\rho}^2 x_n$, il s'ensuit que pour tous les numéros k pour lesquels $\alpha_k^{(n)} \neq 0$ on doit avoir l'égalité

$$(1 - \tau_{n+2}\lambda_k)(1 - \tau_{n+1}\lambda_k) = \bar{\rho}^2.$$

Il en résulte que dans le développement (26) il y a des fonctions propres correspondant seulement à deux valeurs propres différentes (chacune pouvant être un multiple). Posons que c'est λ_i et λ_j. Alors λ_i et λ_j sont des racines de l'équation

$$(1 - \tau_{n+2}\lambda)(1 - \tau_{n+1}\lambda) = \bar{\rho}^2. \qquad (27)$$

Connaissant τ_{n+1}, τ_{n+2} et $\bar{\rho}$, on est en mesure de déduire de cette équation les valeurs propres λ_i et λ_j.

Sans traîner sur les détails, notons que si dans le développement de l'erreur initiale x_0 il existe des fonctions propres correspondant à la valeur propre minimale λ_1 de l'opérateur C et à la valeur propre maximale λ_N, alors, si la suite ρ_k tend vers une valeur limite, dans le développement (26) on ne sera en présence que de ces fonctions propres. Aussi, en résolvant l'équation (27), pourra-t-on trouver λ_1 et λ_N.

L'aboutissement de la suite $\{\rho_k\}$ à une valeur limite avec n fini constitue un cas particulier. Dans le cas général on ne peut qu'affirmer que pour des n suffisamment grands on aura $\rho_{n+1} \approx \rho_{n+2}$ et $x_{n+2} \approx \rho_{n+2}\rho_{n+1}x_n$.

Cette égalité approximative autorise de prévoir que pour un n suffisamment grand les racines de l'équation

$$(1 - \tau_{n+2}\lambda)(1 - \tau_{n+1}\lambda) = \rho_{n+2}\rho_{n+1} \qquad (28)$$

constitueront des approximations suffisamment bonnes de λ_1 et λ_N et, partant, de γ_1 et γ_2 des inégalités (3).

Décrivons cette méthode d'obtention des valeurs approchées de γ_1 et γ_2. Suivant le schéma itératif (2) avec $f = 0$ on procède à $n + 2$ itérations avec les paramètres τ_{k+1} définis dans (12). Comme pour $f = 0$ la solution de (1) est zéro ($u = 0$), on a $z_k = y_k$ et, par suite, ρ_{k+1} peut être obtenu par la formule

$$\rho_{k+1} = \frac{\| z_{k+1} \|_D}{\| z_k \|_D} = \frac{\| y_{k+1} \|_D}{\| y_k \|_D}.$$

Après avoir calculé τ_{n+1}, τ_{n+2}, ρ_{n+1} et ρ_{n+2}, on résout l'équation (28). Les racines de cette équation sont des approximations de γ_1 par le haut et de γ_2 par le bas.

On fournira au § 5 un exemple illustrant la méthode proposée d'obtention de γ_1 et γ_2.

§ 2. Exemples de méthodes du gradient à deux couches

1. Méthode de la plus grande pente. Au § 1 on a étudié les propriétés générales des méthodes itératives à deux couches du type variationnel utilisées pour la recherche de la solution approchée de l'équation linéaire opératorielle

$$Au = f \qquad (1)$$

à opérateur A non dégénéré. Les approximations itératives se calculent suivant le schéma à deux couches

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}}+Ay_k=f, \quad k=0, 1, \ldots, \quad x_0 \in H, \tag{2}$$

tandis que les paramètres d'itération τ_k s'obtiennent suivant la formule

$$\tau_{k+1}=\frac{(Dw_k, z_k)}{(Dw_k, w_k)}, \quad k=0, 1, \ldots, \tag{3}$$

où $w_k = B^{-1}r_k$ est la correction, $r_k = Ay_k - f$ le résidu et $z_k = y_k - u$ l'erreur. Le choix du paramètre τ_{k+1} suivant la formule (3) garantit un minimum à la norme d'erreur z_{k+1} dans H_D avec le passage de y_k à y_{k+1}.

Voyons maintenant les cas particuliers des méthodes du gradient à deux couches. Chaque méthode concrète est fonction du choix de l'opérateur D et possède son domaine d'application. L'opérateur D sera choisi de façon que dans la formule (3) n'apparaissent pour le paramètre d'itération τ_{k+1} que des grandeurs connues au cours du processus d'itérations.

Commençons l'étude des exemples par la méthode de la plus grande pente. Cette méthode ne peut être appliquée qu'au cas d'un opérateur A autoadjoint et défini positif.

Soit l'opérateur A autoadjoint et défini positif dans H. La *méthode de la plus grande pente* se caractérise par le choix suivant de l'opérateur D: $D = A$. L'opérateur B doit être défini positif dans H. Compte tenu des relations $Az_k = Ay_k - f = r_k$ et $A = A^*$, on obtient, à partir de (3), la formule pour le paramètre d'itération τ_{k+1} de la méthode implicite de la plus grande pente

$$\tau_{k+1}=\frac{(r_k, w_k)}{(Aw_k, w_k)}, \quad k=0, 1, \ldots$$

Pour le cas d'un schéma à deux couches explicite (2) ($B = E$) on a $w_k = B^{-1}r_k = r_k$ et la formule pour τ_{k+1} prend la forme

$$\tau_{k+1}=\frac{(r_k, r_k)}{(Ar_k, r_k)}, \quad k=0, 1, \ldots$$

Dans la méthode de la plus grande pente on minimise la norme de l'erreur z_{k+1} dans l'espace énergétique H_A: $\| z_k \|_A = (Az_k, z_k)^{1/2}$. Les conditions de convergence de la méthode ont été formulées dans le théorème 1, duquel on tire les estimations

$$\| z_n \|_A \leqslant \rho^n \| z_0 \|_A, \quad n \geqslant n_0(\varepsilon) = \ln \varepsilon / \ln \rho .$$

La valeur de la quantité ρ est déterminée par les propriétés des opérateurs A et B et par le volume de l'information à priori sur ces derniers. Notons que l'exigence pour l'opérateur $DB^{-1}A = AB^{-1}A$ d'être autoadjoint est équivalente pour la méthode donnée à l'exigence pour B d'être autoadjoint. Donc

1) si $B = B^*$ et les conditions (3), § 1, ou les conditions qui leur sont équivalentes sont remplies (voir ch. VI, § 2, point 3)

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \quad \gamma_1 > 0,$$

on a alors

$$\rho = (1 - \xi)/(1 + \xi), \quad \xi = \gamma_1/\gamma_2;$$

2) si $B \neq B^*$ et les conditions (4), § 1, ou les conditions qui leur sont équivalentes sont remplies (voir ch. VI, § 4, point 2)

$$\gamma_1 (Bx, A^{-1}Bx) \leqslant (Bx, x), \quad (Ax, x) \leqslant \gamma_2 (Bx, x), \quad \gamma_1 > 0,$$

on a alors

$$\rho = \sqrt{1 - \xi}, \quad \xi = \gamma_1/\gamma_2.$$

Notons que si $B = B^*$, la méthode de la plus grande pente possède la propriété asymptotique.

2. Méthode des moindres résidus. Cette méthode peut être utilisée au cas de tout opérateur A non autoadjoint et non dégénéré. Les opérateurs A et B ne sont pas supposés définis positifs isolément, seul l'opérateur B^*A doit être défini positif. La *méthode des moindres résidus* se définit par le choix suivant de l'opérateur D : $D = A^*A$.

La formule (3) pour le paramètre d'itération τ_{k+1} prend dans la méthode des moindres résidus la forme

$$\tau_{k+1} = \frac{(Aw_k, r_k)}{(Aw_k, Aw_k)}, \quad k = 0, 1, \ldots$$

Au cas d'un schéma explicite (2) ($B = E$) il faut que l'opérateur A soit défini positif, tandis que la formule pour τ_{k+1} a la forme

$$\tau_{k+1} = \frac{(Ar_k, r_k)}{(Ar_k, Ar_k)}, \quad k = 0, 1, \ldots$$

L'appellation de la méthode est liée au fait qu'on y minimise la norme du résidu. En effet, on a pour l'opérateur D mentionné

$$\| z_k \|_D^2 = (Dz_k, z_k) = (A^*Az_k, z_k) = \| Az_k \|^2 = \| r_k \|^2.$$

Donc, pour la méthode étudiée la norme d'erreur dans H_D est égale à la norme du résidu qui peut être calculée au cours des itérations puis utilisée pour le contrôle de la fin des itérations.

Du théorème 1 découlent les estimations de la convergence de la méthode

$$\| r_n \| \leqslant \rho^n \| r_0 \|, \quad n \geqslant n_0 (\varepsilon) = \ln \varepsilon / \ln \rho.$$

L'opérateur $DB^{-1}A = A^*AB^{-1}A$ sera autoadjoint dans H si l'opérateur AB^{-1} est autoadjoint, ce qui équivaut à l'exigence pour l'opérateur B^*A d'être autoadjoint. Si cette exigence est remplie, il s'ensuit des conditions (3), § 1, qui, dans le cas considéré, prennent la forme

$$\gamma_1 (Ay, Ay) \leqslant (AB^{-1}Ay, Ay) \leqslant \gamma_2 (Ay, Ay), \quad \gamma_1 > 0,$$

ou après substitution $y = A^{-1}Bx$

$$\gamma_1 (Bx,\ Bx) \leqslant (Ax,\ Bx) \leqslant \gamma_2 (Bx,\ Bx), \quad \gamma_1 > 0. \qquad (4)$$

et la prise en compte du théorème 1, que

$$\rho = (1 - \xi)/(1 + \xi), \quad \xi = \gamma_1/\gamma_2.$$

Notons que la condition $\gamma_1 > 0$ sera remplie si l'exigence susmentionnée, imposée à l'opérateur B^*A d'être autoadjoint est également remplie. Les conditions imposées à l'opérateur B^*A d'être autoadjoint et défini positif seront, par exemple, remplies si l'on admet les hypothèses suivantes: $A = A^* > 0$, $B = B^* > 0$, $AB = BA$.

Dans ce cas les inégalités (4) sont équivalentes à des inégalités plus simples. En effet, en posant dans (4) $x = B^{-1/2}y$ et en utilisant la permutabilité des opérateurs A et B, on obtient

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \quad \gamma_1 > 0. \qquad (5)$$

Les conditions imposant à l'opérateur B^*A d'être autoadjoint et défini positif seront également automatiquement remplies si l'opérateur B est de la forme $B = (A^*)^{-1}B_0$, où B_0 est un opérateur autoadjoint et défini positif. Dans ce cas au lieu des inégalités (5) il faut recourir aux inégalités

$$\gamma_1 B_0 \leqslant A^*A \leqslant \gamma_2 B_0, \quad \gamma_1 > 0, \qquad (6)$$

tandis que dans la formule pour le paramètre τ_{k+1} la correction w_k s'obtiendra de l'équation $B_0 w_k = A^* r_k$.

Si l'opérateur B^*A n'est pas autoadjoint dans H, des conditions (4), § 1, ou des conditions qui leur sont équivalentes

$$\gamma_1 (Bx,\ Bx) \leqslant (Ax,\ Bx), \quad (Ax,\ Ax) \leqslant \gamma_2 (Ax,\ Bx), \quad \gamma_1 > 0$$

et du théorème 1 il s'ensuit que $\rho = \sqrt{1 - \xi}$, $\xi = \gamma_1/\gamma_2$.

3. Méthode des moindres corrections. Cette méthode peut être appliquée à la résolution de l'équation (1) avec opérateur A non autoadjoint mais défini positif. Il faut que l'opérateur B soit autoadjoint, défini positif et borné. La *méthode des moindres corrections* est définie par le choix suivant de l'opérateur D: $D = A^*B^{-1}A$.

La formule (3) du paramètre d'itération τ_{k+1} prend dans la méthode des moindres corrections la forme

$$\tau_{k+1} = \frac{(Aw_k,\ w_k)}{(B^{-1}Aw_k,\ Aw_k)}, \quad k = 0, 1, \ldots$$

Au cas d'un schéma explicite (2) ($B = E$) les méthodes des moindres corrections et des moindres résidus se confondent.

Dans la méthode des moindres corrections on minimise la norme de correction dans H_B. En effet, pour l'opérateur D choisi on obtient

$$\| z_k \|_D^2 = (Dz_k,\ z_k) = (A^*B^{-1}Az_k,\ z_k) = (w_k,\ r_k) = (Bw_k,\ w_k) = \| w_k \|_B^2.$$

La norme de correction dans H_B peut être calculée au cours des itérations et utilisée pour le contrôle de la fin des itérations.

Du théorème 1 on déduit les estimations de la convergence de la méthode

$$\| w_n \|_B \leqslant \rho^n \| w_0 \|_B, \quad n \geqslant n_0 (\varepsilon) = \ln \varepsilon / \ln \rho.$$

L'opérateur $DB^{-1}A = A^*B^{-1}AB^{-1}A$ est autoadjoint dans H en même temps que l'opérateur A. Aussi:

1) si $A = A^*$ et les conditions (3) du § 1 ou les conditions qui leur sont équivalentes (voir ch. VI, § 2, point 3) sont remplies

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \quad \gamma_1 > 0,$$

on obtient

$$\rho = (1 - \xi)/(1 + \xi), \quad \xi = \gamma_1/\gamma_2;$$

2) si $A \neq A^*$ et les conditions (4) du § 1 ou les conditions qui leur sont équivalentes (voir ch. VI, § 4, point 2) sont remplies

$$\gamma_1 B \leqslant A, \ (Ax, B^{-1}Ax) \leqslant \gamma_2 (Ax, x), \quad \gamma_1 > 0,$$

on a

$$\rho = \sqrt{1 - \xi}, \quad \xi = \gamma_1/\gamma_2.$$

Notons que par comparaison aux méthodes de la plus grande pente et des moindres résidus dans la méthode des moindres corrections l'opérateur B doit être inversé non pas une fois mais deux fois, d'abord pour calculer la correction w_k et, ensuite, pour calculer $B^{-1}Aw_k$.

Notons de même que si $A = A^*$, la méthode des moindres corrections possède la propriété asymptotique.

4. Méthode des moindres erreurs. Cette méthode peut être appliquée, comme celle des moindres résidus, dans le cas de tout opérateur A non autoadjoint et non dégénéré. La *méthode des moindres erreurs* se définit par le choix suivant des opérateurs B et D:

$$B = (A^*)^{-1} B_0, \quad D = B_0,$$

où B_0 est l'opérateur autoadjoint et défini positif dans H.

En portant dans la formule (3) du paramètre d'itération τ_{k+1} l'opérateur D choisi et compte tenu de ce que $w_k = B^{-1} r_k = B_0^{-1} A^* r_k$, on obtient la formule pour τ_{k+1} de la méthode des moindres erreurs

$$\tau_{k+1} = \frac{(r_k, r_k)}{(Aw_k, r_k)}, \quad k = 0, 1, \ldots$$

La correction w_k s'obtient de l'équation $B_0 w_k = A^* r_k$.

Au cas d'un schéma explicite ($B_0 = E$) la formule pour τ_{k+1} prend la forme

$$\tau_{k+1} = \frac{(r_k, r_k)}{(A^* r_k, A^* r_k)}, \quad k = 0, 1, \ldots$$

Dans la méthode des moindres erreurs on minimise la norme d'erreur dans H_{B_0}. Dans cette méthode l'opérateur $DB^{-1}A = A^*A$

est autoadjoint dans H, tandis que les conditions (3) du § 1 prennent la forme des inégalités (6). Il s'ensuit du théorème 1 l'estimation sur la convergence de la méthode

$$\| z_n \|_{B_0} \leqslant \rho^n \| z_0 \|_{B_0}, \qquad n \geqslant n_0 (\varepsilon) = \ln \varepsilon / \ln \rho,$$

où $\rho = (1 - \xi)/(1 + \xi)$, $\xi = \gamma_1/\gamma_2$, quant à γ_1 et γ_2, ils sont définis dans (6).

La méthode des moindres erreurs possède toujours la propriété asymptotique.

5. Exemple d'application des méthodes à deux couches. A titre d'illustration de l'application des méthodes du gradient à deux couches, étudions la résolution du problème modèle par la méthode explicite de la plus grande pente. En guise d'exemple, prenons le problème discret de Dirichlet pour l'équation de Poisson sur maillage carré $\bar{\omega} = \{x_{ij} = (ih, jh),\ 0 \leqslant i \leqslant N,\ 0 \leqslant j \leqslant N,\ h = 1/N\}$ dans un carré unitaire

$$\Lambda u = u_{\bar{x}_1 x_1} + u_{\bar{x}_2 x_2} = -\varphi, \quad x \in \omega, \quad u|_\gamma = g. \tag{7}$$

Introduisons l'espace H composé des fonctions de mailles données sur ω avec produit scalaire $(u, v) = \sum\limits_{x \in \omega} u(x)\, v(x)\, h^2$.

L'opérateur A sur H est défini de la façon suivante: $Ay = -\Lambda v$, $y \in H$, où $v(x) = y(x)$ pour $x \in \omega$ et $v|_\gamma = 0$. Ecrivons le problème (7) sous forme d'une équation opératorielle

$$Au = f, \tag{8}$$

où f ne diffère de φ que dans les nœuds voisins de la frontière

$$f = \varphi + \frac{\varphi_1}{h^2} + \frac{\varphi_2}{h^2},$$

$$\varphi_1 = \begin{cases} g(0, x_2), & x_1 = h, \\ 0, & 2h \leqslant x_1 \leqslant 1 - 2h, \\ g(1, x_2), & x_1 = 1 - h, \end{cases}$$

$$\varphi_2 = \begin{cases} g(x_1, 0), & x_2 = h, \\ 0, & 2h \leqslant x_2 \leqslant 1 - 2h, \\ g(x_1, 1), & x_2 = 1 - h. \end{cases}$$

L'opérateur A est autoadjoint et défini positif dans H. Aussi, pour résoudre l'équation (8), peut-on appliquer la méthode de la plus grande pente. Le schéma itératif explicite est de la forme

$$\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f \quad \text{ou} \quad y_{k+1} = y_k - \tau_{k+1} r_k, \quad k = 0, 1, \ldots,$$

quant aux paramètres d'itération τ_k, on les obtient par la formule

$$\tau_{k+1}=\frac{(r_k,\ r_k)}{(Ar_k,\ r_k)},\quad r_k=Ay_k-f,\quad k=0,\ 1,\ \ldots$$

Donnons les formules de calcul et calculons le nombre d'opérations arithmétiques que coûte une itération.

Compte tenu de la définition de l'opérateur A et du second membre f, on peut écrire les formules de calcul sous la forme suivante:

1) $r_k(x_{ij})=-(y_k)_{\bar{x}_1x_1}-(y_k)_{\bar{x}_2x_2}-\varphi(x_{ij}),\quad 1\leqslant i,\ j\leqslant N-1,$

$y_k|_\gamma=g;$

2) $(r_k,\ r_k)=\sum\limits_{i=1}^{N-1}\sum\limits_{j=1}^{N-1} r_k^2(x_{ij})\,h^2,$

$$(Ar_k,\ r_k)=-\sum_{i=1}^{N-1}\sum_{j=1}^{N-1} r_k(x_{ij})\,[(r_k)_{\bar{x}_1x_1}+(r_k)_{\bar{x}_2x_2}]\,h^2,\quad r_k|_\gamma=0,$$

$$\tau_{k+1}=\frac{(r_k,\ r_k)}{(Ar_k,\ r_k)};$$

3) $y_{k+1}(x_{ij})=y_k(x_{ij})-\tau_{k+1}r_k(x_{ij}),\quad 1\leqslant i,\ j\leqslant N-1.$

L'approximation initiale y_0 est une fonction de maille arbitraire dans ω qui prend sur γ les valeurs données $y_0|_\gamma=g$.

Calculons le nombre d'opérations arithmétiques. Si le calcul des différences divisées s'effectue suivant la formule

$$u_{\bar{x}_1x_1}+u_{\bar{x}_2x_2}=\frac{1}{h^2}(u_{i+1,\ j}+u_{i-1,\ j}+u_{i,\ j+1}+u_{i,\ j-1}-4u_{ij}),$$

il faudra, pour le calcul de r_k, 6 $(N-1)^2$ additions et 2 $(N-1)^2$ multiplications et divisions. Pour le calcul de $(r_k,\ r_k)$ il faut $(N-1)^2$ additions et $(N-1)^2$ multiplications, pour $(Ar_k,\ r_k)$ 6 $(N-1)^2$ additions et 2 $(N-1)^2$ multiplications, pour y_{k+1} $(N-1)^2$ additions et $(N-1)^2$ multiplications. En tout, on aura besoin de 14 $(N-1)^2$ additions et 6 $(N-1)^2$ multiplications et divisions. Exactement la moitié de ce nombre total d'opérations sera dépensée au calcul des produits scalaires, c'est-à-dire au calcul du paramètre d'itération τ_{k+1}. Par conséquent, une seule opération d'itération de la méthode de la plus grande pente est environ deux fois plus laborieuse qu'une seule opération d'itération de la méthode itérative simple ou de la méthode de Tchébychev, où les paramètres τ_{k+1} sont connus à priori. Pour les méthodes implicites, cette différence sera moindre, vu que le calcul des produits scalaires exigera le même nombre d'opérations que la méthode explicite, tandis qu'au nombre total d'opérations s'ajouteront les opérations arithmétiques impliquées par l'inversion de l'opérateur B.

Calculons maintenant le nombre total d'opérations arithmétiques $Q(\varepsilon)$ qu'il faut accomplir pour obtenir la précision relative ε. Il faut pour cela apprécier le nombre d'itérations $n_0(\varepsilon)$. Au point 1 on a obtenu l'estimation suivante:

$$n_0(\varepsilon)=\frac{\ln\varepsilon}{\ln\rho},\quad \rho=\frac{1-\xi}{1+\xi},\quad \xi=\frac{\gamma_1}{\gamma_2},$$

où γ_1 et γ_2, au cas d'un schéma explicite, sont les bornes de l'opérateur A: $\gamma_1 E\leqslant A\leqslant\gamma_2 E$.

Pour l'exemple étudié γ_1 et γ_2 coïncident avec les valeurs propres minimale δ et maximale Δ de l'opérateur de différences de Laplace Λ. On sait que

$$\delta=\frac{8}{h^2}\sin^2\frac{\pi h}{2},\quad \Delta=\frac{8}{h^2}\cos^2\frac{\pi h}{2}.$$

Par conséquent,

$$\rho=\frac{1-\xi}{1+\xi}=1-2\sin^2\frac{\pi h}{2},\quad \xi=\frac{\delta}{\Delta}=\operatorname{tg}^2\frac{\pi h}{2},$$

et, par suite, si $h\ll 1$, on a

$$n_0(\varepsilon)\approx\frac{2\ln\frac{1}{\varepsilon}}{\pi^2h^2}\approx 0,\ 2N^2\ln\frac{1}{\varepsilon}.$$

Si l'on assimile les opérations d'addition à celles de multiplication et de division, on aura alors besoin pour une seule itération d'environ $20N^2$ opérations. Pour le nombre total d'opérations arithmétiques l'estimation $Q(\varepsilon)\approx 4N^4\ln\frac{1}{\varepsilon}$ sera donc juste.

§ 3. Méthodes des directions conjuguées à trois couches

1. Position du problème sur le choix des paramètres d'itération. Appréciation de la vitesse de convergence. Pour trouver la solution approchée de l'équation linéaire opératorielle

$$Au=f \tag{1}$$

à opérateur A non dégénéré, on a étudié au § 1 les méthodes itératives à deux couches du type variationnel. Le schéma itératif de ces méthodes prend la forme

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}}+Ay_k=f,\quad k=0,\ 1,\ \ldots,\quad y_0\in H, \tag{2}$$

quant aux paramètres d'itération τ_{k+1}, ils sont choisis sur la base de la condition du minimum de la norme d'erreur z_{k+1} dans l'espace énergétique H_D. Rappelons que sur la suite $\{y_k\}$, construite suivant la formule (2), s'effectue la minimisation de proche en proche de la

fonctionnelle $I(y) = (D(y-u), y-u)$ dont le minimum est atteint avec la résolution de l'équation (1), c'est-à-dire pour $y = u$.

Cette stratégie de minimisation locale n'est pas toutefois optimale, car ce qui nous intéresse finalement c'est le minimum global de la fonctionnelle $I(y)$, et, si est donnée une certaine valeur de cette fonctionnelle, on doit aboutir au minimum cherché par un nombre minimal d'itérations. Or la minimisation locale à chaque itération conduit à la solution de ce problème par une voie qui n'est pas la plus courte.

Il est tout naturel de tenter de choisir aussitôt les paramètres τ_k sur la base de la condition du minimum de la norme d'erreur z_n dans H_D en n pas, c'est-à-dire au cours du passage de y_0 à y_n. On s'est déjà heurté à une situation analogue au chapitre VI lors de l'étude de la méthode de Tchébychev et de la méthode itérative simple. Il s'est alors avéré que la méthode qui converge le plus vite est celle dont les paramètres d'itération sont choisis sur la base de la condition du minimum de la norme de l'opérateur résolvant et non pas de l'opérateur de transfert d'une itération à l'autre. Cette propriété s'observe pour les méthodes itératives du type variationnel. On montrera que les méthodes itératives étudiées dans ce paragraphe, et dont les paramètres τ_k sont choisis sur la base de la condition mentionnée plus haut, convergent beaucoup plus vite que les méthodes du gradient à deux couches. En outre, au cas d'un espace de dimension finie H ces méthodes deviennent des méthodes à itérations finies pour toute approximation initiale, autrement dit la solution exacte de l'équation (1) peut être obtenue au bout d'un nombre fini d'itérations.

Passons à la construction de la *méthode des directions conjuguées*. On supposera que l'opérateur $DB^{-1}A$ est autoadjoint et défini positif dans H. Effectuons suivant le schéma (2) n itérations. En passant du problème sur l'erreur $z_k = y_k - u$ au problème pour $x_k = D^{1/2}z_k$, on obtient comme auparavant

$$x_{k+1} = S_{k+1}x_k, \quad k = 0, 1, \ldots, n-1,$$
$$S_k = E - \tau_k C, \quad C = D^{1/2}B^{-1}AD^{-1/2}.$$

De là, il vient

$$x_n = T_n x_0, \quad T_n = \prod_{j=1}^{n} (E - \tau_j C). \tag{3}$$

L'opérateur résolvant T_n constitue un polynôme opératoriel de degré n relativement à l'opérateur C avec coefficients dépendant des paramètres $\tau_1, \tau_2, \ldots, \tau_n$

$$T_n = P_n(C) = E + \sum_{j=1}^{n} a_j^{(n)} C^j, \quad a_n^{(n)} \neq 0. \tag{4}$$

En vertu de l'égalité $\| x_n \| = \| z_n \|_D$ le problème du choix des paramètres d'itération τ_k, posé plus haut, se formule de la façon suivante : parmi tous les polynômes de la forme (4) il faut choisir celui pour lequel la norme $x_n = P_n(C)\, x_0$ est minimale, autrement dit il faut choisir les coefficients $a_1^{(n)}, a_2^{(n)}, \ldots, a_n^{(n)}$ du polynôme $P_n(C)$ sur la base de la condition du minimum de la norme x_n dans H.

Ce problème sera résolu au point suivant, mais d'abord apprécions la vitesse de convergence de la méthode des directions conjuguées construite sur la base du principe formulé plus haut du choix des paramètres. L'estimation sera obtenue en utilisant l'information à priori sur les opérateurs du schéma sous forme de γ_1 et γ_2 constituant des constantes de l'équivalence énergétique des opérateurs autoadjoints D et $DB^{-1}A$:

$$\gamma_1 D \leqslant DB^{-1}A \leqslant \gamma_2 D, \quad \gamma_1 > 0, \quad DB^{-1}A = (DB^{-1}A)^*. \tag{5}$$

Soit $P_n(C)$ le polynôme cherché. Alors de (3), (4) s'ensuit l'estimation pour x_n :

$$\| x_n \| = \| P_n(C)\, x_0 \| = \min_{\{Q_n\}} \| Q_n(C)\, x_0 \| \leqslant \min_{\{Q_n\}} \| Q_n(C) \| \, \| x_0 \|,$$

où le minimum est recherché parmi les polynômes $Q_n(C)$ normés en vertu de (4) par la condition $Q_n(0) = E$.

Apprécions le minimum de la norme du polynôme $Q_n(C)$. Il s'ensuit de (5) que l'opérateur $C = D^{-1/2}(DB^{-1}A)\, D^{-1/2}$ est autoadjoint dans H, tandis que γ_1 et γ_2 sont ses bornes : $C = C^*$, $\gamma_1 E \leqslant$ $\leqslant C \leqslant \gamma_2 E$, $\gamma_1 > 0$. On a donc l'estimation

$$\min_{\{Q_n\}} \| Q_n(C) \| \leqslant \min_{\{Q_n\}} \max_{\gamma_1 \leqslant t \leqslant \gamma_2} | Q_n(t) |.$$

Il s'ensuit finalement du § 2, ch. VI, que le problème de construction du polynôme normé par la condition $Q_n(0) = 1$ et dont le maximum du module sur le tronçon $[\gamma_1, \gamma_2]$ est minimal, se résout au moyen du polynôme de Tchébychev de première espèce pour lequel

$$\max_{\gamma_1 \leqslant t \leqslant \gamma_2} | Q_n(t) | = q_n, \quad q_n = \frac{2\rho_1^n}{1+\rho_1^{2n}}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

On a donc pour x_n l'estimation $\| x_n \| \leqslant q_n \| x_0 \|$.

On a ainsi démontré le

Théorème 2. *Si les conditions* (5) *sont remplies, la méthode itérative des directions conjuguées converge dans* H_D, *et on a pour l'erreur* z_n *avec tout* n *l'estimation* $\| z_n \|_D \leqslant q_n \| z_0 \|_D$. *En outre, l'estimation du nombre d'itérations prend la forme*

$$n \geqslant n_0(\varepsilon) = \ln(0.5\varepsilon)/\ln \rho_1,$$

où $\rho_1 = (1-\sqrt{\xi})/(1+\sqrt{\xi})$, $\xi = \gamma_1/\gamma_2$.

2. Formules des paramètres d'itération. Schéma itératif à trois couches. Abordons maintenant la construction du polynôme $P_n\ (C)$. En utilisant (3) et (4), on calcule la norme x_n:

$$\| x_n \|^2 = (P_n (C)\, x_0,\quad P_n (C)\, x_0) =$$

$$= \| x_0 \|^2 + 2 \sum_{j=1}^{n} a_j^{(n)} (C^j x_0,\ x_0) + \sum_{j=1}^{n} \sum_{i=1}^{n} a_j^{(n)} a_i^{(n)} (C^j x_0,\ C^i x_0).$$

La norme x_n est une fonction des paramètres $a_1^{(n)}, a_2^{(n)}, \ldots, a_n^{(n)}$. En égalant à zéro les dérivées partielles de $\| x_n \|^2$ en $a_j^{(n)}$

$$\frac{\partial \| x_n \|^2}{\partial a_j^{(n)}} = 2 \sum_{i=1}^{n} a_i^{(n)} (C^j x_0,\ C^i x_0) + 2 (C^j x_0,\ x_0),\quad j = 1,\ 2,\ \ldots,\ n,$$

on obtient le système d'équations algébriques linéaires

$$\sum_{i=1}^{n} a_i^{(n)} (C^j x_0,\ C^i x_0) + (C^j x_0,\ x_0) = 0,\quad j = 1,\ 2,\ \ldots,\ n. \qquad (6)$$

Pour l'opérateur C autoadjoint et défini positif dans H le système (6) fournit les conditions du minimum de la norme x_n dans H.

Le problème de la construction du polynôme optimal $P_n\ (C)$ est donc en principe résolu. Les coefficients du polynôme $a_1^{(n)}, a_2^{(n)}, \ldots$ $\ldots, a_n^{(n)}$ seront obtenus par résolution du système (6). Mais d'abord construisons les formules pour le calcul de l'approximation itérative y_n. La première voie consiste à se servir du schéma itératif (2). Mais il faudra alors trouver les racines du polynôme $P_n\ (t)$ et, ensuite, prendre en guise de τ_k les valeurs inverses des racines. Ce procédé n'est pas économique.

La seconde voie consiste à utiliser pour le calcul de y_n les coefficients du polynôme. De (3), (4) et après la substitution $x_k = D^{1/2} z_k$, où $z_k = y_k - u$, il vient

$$y_n - u = D^{-1/2} P_n\ (C)\ D^{1/2}\ (y_0 - u). \qquad (7)$$

Utilisant (4) et l'égalité $D^{-1/2} C^j D^{1/2} = (B^{-1}A)^j$, on obtient

$$D^{-1/2} P_n (C)\, D^{1/2} = E + \sum_{j=1}^{n} a_j^{(n)} (B^{-1}A)^j.$$

En portant cette égalité dans (7), on trouve

$$y_n = y_0 + \sum_{j=1}^{n} a_j^{(n)} (B^{-1}A)^j (y_0 - u) = y_0 + \sum_{j=1}^{n} a_j^{(n)} (B^{-1}A)^{j-1} w_0, \qquad (8)$$

où w_0 est la correction, $w_0 = B^{-1}A\ (y_0 - u) = B^{-1} r_0$, $r_0 = A y_0 - f$.

Ce procédé n'est pas également optimal. Pour chaque nouveau n il faut recommencer à résoudre le système (6).

On va maintenant montrer que la suite $y_1, y_2, \ldots, y_k, \ldots$, construite en conformité avec (6), (8) pour $n = 1, 2, \ldots$, peut être obtenue à partir du schéma à trois couches suivant :

$$By_{k+1} = \alpha_{k+1} (B - \tau_{k+1}A) y_k + (1 - \alpha_{k+1}) By_{k-1} + \alpha_{k+1}\tau_{k+1}f,$$
$$k = 1, 2, \ldots, \quad (9)$$
$$By_1 = (B - \tau_1 A) y_0 + \tau_1 f, \quad y_0 \in H.$$

Il faut pour cela indiquer le jeu de paramètres $\{\tau_k\}$ et $\{\alpha_k\}$ pour lequel la norme de l'erreur équivalente x_k soit minimale pour tout k. En effet, de l'équation pour l'erreur x_k au cas du schéma (9)

$$x_{k+1} = \alpha_{k+1} (E - \tau_{k+1}C) x_k + (1 - \alpha_{k+1}) x_{k-1}, \quad k = 1, 2, \ldots, \quad (10)$$
$$x_1 = (E - \tau_1 C) x_0,$$

on tire que $x_k = P_k (C) x_0$, où le polynôme $P_k (C)$ a la forme (4) ($n = k$). Aussi si les paramètres $\{\tau_k\}$ et $\{\alpha_k\}$ seront choisis dans (9) de manière que pour tout $n = 1, 2, \ldots$ les conditions (6) demeurent remplies, alors, les approximations itératives y_n, construites suivant (9), coïncideront avec les approximations obtenues suivant les formules (6), (8) pour tout n.

Construisons le jeu cherché des paramètres $\{\tau_k\}$ et $\{\alpha_k\}$. Pour cela énonçons le lemme suivant.

L e m m e. *Les conditions nécessaires et suffisantes du minimum de la norme x_n dans H pour tout $n \geqslant 1$ sont*

$$(Cx_j, x_n) = 0, \quad j = 0, 1, \ldots, n - 1. \quad (11)$$

En effet, de (4), (6) il s'ensuit que les conditions (6), qui sont les conditions du minimum de la norme x_n, sont équivalentes aux suivantes :

$$(C^j x_0, x_n) = 0, \quad j = 1, 2, \ldots, n, \quad (12)$$

pour tout $n = 1, 2, \ldots$ De là on obtient pour $j \leqslant n - 1$

$$(Cx_0, x_n) + \sum_{i=2}^{j+1} a_{i-1}^{(j)} (C^i x_0, x_n) = (Cx_j, x_n) = 0,$$

autrement dit les conditions (11) sont nécessaires.

Démontrons maintenant que les conditions (11) sont suffisantes. Supposons que les conditions (11) sont remplies. Montrons qu'alors sont également remplies les conditions (12). De (11) pour $j = 0$ on tire que les égalités (12) se vérifient pour $j = 1$. L'exactitude de (12) pour $j \geqslant 2$ sera démontrée par induction. Supposons que pour $j \leqslant k$ les conditions (12) sont remplies, c'est-à-dire que $(C^j x_0, x_n) = 0$, $j = 1, 2, \ldots, k$. Montrons qu'elles sont également satisfaites pour $j = k + 1$ au cas où les conditions (11) sont remplies.

En effet, de (11) pour $j = k$, on obtient

$$0 = (Cx_k, \ x_n) = (CP_k(C)\,x_0, \ x_n) =$$
$$= (Cx_0, \ x_n) + \sum_{j=1}^{k} a_j^{(k)} (C^{j+1}x_0, \ x_n) = a_k^{(k)} (C^{k+1}x_0, \ x_n).$$

Par conséquent, $(C^{k+1}x_0, x_n) = 0$. Le lemme est démontré.

Profitons maintenant de ce lemme pour la construction du jeu des paramètres $\{\tau_k\}$ et $\{\alpha_k\}$ pour le schéma (9). Pour abréger les calculs, admettons que y_1 dans le schéma (9) s'obtient avec la formule générale (9) pour $\alpha_1 = 1$.

Analysons le schéma (10). x_1 s'obtenant suivant le schéma à deux couches, on aboutit, sur la base du § 1, au choix optimal du paramètre τ_1 à l'aide de la formule

$$\tau_1 = \frac{(Cx_0, \ x_0)}{(Cx_0, \ Cx_0)}.$$

La construction des paramètres $\tau_2, \tau_3, \ldots$ et $\alpha_2, \alpha_3, \ldots$ sera réalisée progressivement. Admettons que les paramètres d'itération $\tau_1, \tau_2, \ldots, \tau_k$ et $\alpha_1, \alpha_2, \ldots, \alpha_k$ ont été choisis de façon optimale. Vu que ces paramètres déterminent les approximations $y_1, y_2, \ldots, y_k$, il découle du lemme que les conditions

$$(Cx_j, \ x_i) = 0, \quad j = 0, 1, \ldots, i-1, \quad i = 1, 2, \ldots, k \qquad (13)$$

sont remplies.

Choisissons maintenant les paramètres τ_{k+1} et α_{k+1} définissant l'approximation y_{k+1}. Il s'ensuit du lemme que la norme x_{k+1} sera minimale si sont remplies les conditions

$$(Cx_j, \ x_{k+1}) = 0, \quad j = 0, 1, \ldots, k. \qquad (14)$$

A partir de ces conditions cherchons les paramètres τ_{k+1} et α_{k+1}. Montrons d'abord que de (13) il s'ensuit que les conditions (14) sont remplies pour $j \leqslant k-2$, et. ensuite, des deux conditions restantes de (14) pour $j = k-1$ et $j = k$, on obtient les formules pour τ_{k+1} et α_{k+1}.

Bref, soit $j \leqslant k-2$. De (10) et (13) il vient

$$(x_{k+1}, \ Cx_j) = \alpha_{k+1}(x_k, \ Cx_j) - \alpha_{k+1}\tau_{k+1}(Cx_k, \ Cx_j) +$$
$$+ (1-\alpha_{k+1})(x_{k-1}, \ Cx_j) = -\alpha_{k+1}\tau_{k+1}(Cx_k, \ Cx_j).$$

Montrons que $(Cx_k, Cx_j) = 0$ pour $j \leqslant k-2$. En effet, de (10) pour $k = j$, on obtient

$$Cx_j = \frac{1}{\tau_{j+1}} x_j - \frac{1}{\tau_{j+1}\alpha_{j+1}} [x_{j+1} - (1-\alpha_{j+1})\,x_{j-1}], \quad j \geqslant 0. \qquad (15)$$

En utilisant le fait que l'opérateur C est autoadjoint, de même que les conditions (13), on obtient de ce qui précède pour $j \leqslant k-2$

$$(Cx_k, Cx_j) = = \frac{1}{\tau_{j+1}}(Cx_j, x_k) - \frac{1}{\tau_{j+1}\alpha_{j+1}}[(Cx_{j+1}, x_k) - (1-\alpha_{j+1})(Cx_{j-1}, x_k)] = 0.$$

Par conséquent, $(x_{k+1}, Cx_j) = 0$ pour $j \leqslant k-2$.

Cherchons maintenant τ_{k+1} et α_{k+1}. En posant dans (14) $j = k-1$ et $j = k$, on obtient de (10) et (13)

$$0 = (Cx_{k-1}, x_{k+1}) = -\alpha_{k+1}\tau_{k+1}(Cx_k, Cx_{k-1}) + (1-\alpha_{k+1})(Cx_{k-1}, x_{k-1}), \quad (16)$$

$$0 = (Cx_k, x_{k+1}) = \alpha_{k+1}[(Cx_k, x_k) - \tau_{k+1}(Cx_k, Cx_k)].$$

De la seconde équation on tire aussitôt le paramètre τ_{k+1} :

$$\tau_{k+1} = \frac{(Cx_k, x_k)}{(Cx_k, Cx_k)}. \quad (17)$$

De la première équation on élimine l'expression (Cx_k, Cx_{k-1}). Posons pour cela dans (15) $j = k-1$ et multiplions scalairement le premier et le second membres de (15) par Cx_k.

Puisque l'opérateur C est autoadjoint de la condition (13), on obtient

$$(Cx_k, Cx_{k-1}) = \frac{1}{\tau_k}(Cx_{k-1}, x_k) - \frac{1}{\tau_k\alpha_k}(Cx_k, x_k) + \frac{1-\alpha_k}{\tau_k\alpha_k}(Cx_{k-2}, x_k) = -\frac{1}{\tau_k\alpha_k}(Cx_k, x_k).$$

En portant cette expression dans (16), il vient

$$\frac{\alpha_{k+1}\tau_{k+1}}{\alpha_k\tau_k}\frac{(Cx_k, x_k)}{(Cx_{k-1}, x_{k-1})} + (1-\alpha_{k+1}) = 0.$$

De cette égalité on obtient la formule de récurrence pour le paramètre α_{k+1} :

$$\alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k}\frac{(Cx_k, x_k)}{(Cx_{k-1}, x_{k-1})}\frac{1}{\alpha_k}\right)^{-1}. \quad (18)$$

Bref, en admettant que les paramètres d'itération $\tau_1, \tau_2, \ldots, \tau_k$ et $\alpha_1, \alpha_2, \ldots, \alpha_k$ ont déjà été choisis, on aboutit à des formules pour les paramètres τ_{k+1} et α_{k+1}. Comme $\alpha_1 = 1$ et $\tau_1 = \dfrac{(Cx_0, x_0)}{(Cx_0, Cx_0)}$, les formules (17), (18) définissent donc les paramètres τ_{k+1} et α_{k+1} pour tout k.

En portant $x_k = D^{1/2}z_k$ dans (17) et (18) et compte tenu de ce que

$$C = D^{-1/2}(DB^{-1}A)D^{-1/2} \quad \text{et} \quad Az_k = r_k, \quad B^{-1}r_k = w_k,$$

on obtient les formules suivantes pour les paramètres d'itération τ_{k+1} et α_{k+1}:

$$\tau_{k+1} = \frac{(Dw_k, z_k)}{(Dw_k, w_k)}, \quad k = 0, 1, \ldots, \tag{19}$$

$$\alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(Dw_k, z_k)}{(Dw_{k-1}, z_{k-1})} \frac{1}{\alpha_k}\right)^{-1}, \tag{20}$$

$$k = 1, 2, \ldots, \quad \alpha_1 = 1.$$

Ainsi donc la méthode des directions conjuguées est décrite par le schéma à trois couches (9), dont les paramètres d'itération τ_{k+1} et α_{k+1} sont choisis suivant les formules (19), (20). Pour cette méthode se vérifie le théorème 2 démontré auparavant.

Il s'ensuit des formules (19), (20) que les paramètres d'itération τ_{k+1} sont choisis dans la méthode des directions conjuguées et les méthodes du gradient à deux couches suivant les mêmes formules, tandis que pour le calcul des paramètres α_{k+1} il ne faut calculer aucun produit scalaire supplémentaire. Aussi pour le calcul des paramètres d'itération dans les méthodes à deux et trois couches du type variationnel dépense-t-on pratiquement un même nombre d'opérations arithmétiques. Mais en même temps il découle des théorèmes 1 et 2 que les méthodes des directions conjuguées convergent sensiblement plus vite que les méthodes du gradient.

Montrons maintenant que si H est un espace de dimension finie ($H = H_N$), alors les méthodes des directions conjuguées convergent en un nombre fini d'itérations ne dépassant pas les dimensions de l'espace. En effet, il s'ensuit du lemme que pour des erreurs équivalentes x_k de la méthode des directions conjuguées doivent se vérifier les égalités $(Cx_j, x_n) = (x_j, x_n)_C = 0$, $j = 0, 1, \ldots, n$. Donc le système des vecteurs $x_0, x_1, \ldots, x_n$ pour tout n doit être orthogonal dans H_C. Comme dans H_N on ne peut construire plus de N vecteurs orthogonaux, il s'ensuit que $x_N = 0$ et $z_N = y_N - u = 0$. Donc avec la classe d'approximations initiales arbitraires y_0 les méthodes des directions conjuguées convergent en N itérations vers la solution précise de l'équation (1).

Avec des approximations initiales spéciales y_0 ces méthodes convergent en un nombre moindre d'itérations. En effet, supposons que y_0 est tel que dans le développement en x_0 suivant les fonctions propres de l'opérateur C figurent $N_0 < N$ fonctions, c'est-à-dire que x_0 appartient au sous-espace H_{N_0} invariant par rapport à l'opérateur C. Alors il devient évident que tous les $x_k \in H_{N_0}$. Aussi dans ce cas le processus d'itération convergera-t-il en N_0 itérations.

Il ne s'ensuit pas de ce qui vient d'être dit que l'estimation de la convergence de la méthode résultant du théorème 2 est très grossière et que l'égalité $\| z_n \|_D = q_n \| z_0 \|_D$ n'est jamais atteinte. Il est

possible de construire un exemple d'équation (1) et d'indiquer pour tout $n < N$ une telle approximation initiale y_0 pour laquelle l'égalité mentionnée sera vérifiée.

3. Variantes des formules de calculs. Donnons maintenant quelques procédés de mise en œuvre des méthodes des directions conjuguées à trois couches. Sur la base de (9), (19) et (20), formons l'algorithme suivant:

1) d'après y_0 donné on calcule le résidu $r_0 = Ay_0 - f$;

2) on résout l'équation pour la correction $Bw_0 = r_0$;

3) on calcule le paramètre τ_1 suivant la formule (19);

4) on obtient l'approximation y_1 suivant la formule $y_1 = y_0 - \tau_1 w_0$.

Ensuite, pour $k = 1, 2, \ldots$ on effectue successivement les opérations suivantes:

5) on calcule le résidu $r_k = Ay_k - f$ et on résout l'équation de la correction $Bw_k = r_k$;

6) avec les formules (19), (20) on calcule les paramètres τ_{k+1} et α_{k+1};

7) on obtient l'approximation y_{k+1} suivant la formule

$$y_{k+1} = \alpha_{k+1} y_k + (1 - \alpha_{k+1}) y_{k-1} - \alpha_{k+1}\tau_{k+1} w_k.$$

Avec l'algorithme décrit, pour trouver y_{k+1}, on utilise donc y_{k-1}, y_k et w_k qui doivent être mémorisés. On fournira plus loin la forme des formules (19) et (20) pour quelques choix concrets de l'opérateur D. En attendant, on se limitera d'indiquer que ces formules peuvent comporter, outre la quantité mémorisée w_k, le résidu r_k qui n'est pas retenu. Pour le calculer, on peut se servir soit de l'égalité $r_k = Bw_k$, au cas où le calcul de Bw_k n'est pas trop laborieux, soit de la définition du résidu $r_k = Ay_k - f$.

En pratique il existe également d'autres algorithmes de la mise en œuvre de la méthode des directions conjuguées. Indiquons l'un d'eux. A cette fin le schéma (9) sera traité comme un schéma à correction. De (9), il vient

$$y_{k+1} = y_k - a_{k+1}s_k, \quad s_{k+1} = w_{k+1} + b_{k+1}s_k, \quad k = 0, 1, \ldots, s_0 = w_0, \tag{21}$$

où $w_k = B^{-1}r_k$, $r_k = Ay_k - f$, quant aux paramètres a_{k+1} et b_k, ils sont reliés à α_{k+1} et τ_{k+1} par les formules suivantes:

$$a_{k+1} = \alpha_{k+1}\tau_{k+1}, \quad b_k = (\alpha_{k+1} - 1)\,\alpha_k\tau_k/(\alpha_{k+1}\tau_{k+1}).$$

Cherchons les expressions de b_k et a_{k+1}. De (19), (20), il vient

$$b_k = (Dw_k, z_k)/(Dw_{k-1}, z_{k-1}), \quad k = 1, 2, \ldots \tag{22}$$

On obtient sans peine pour a_{k+1} à partir des mêmes formules les relations de récurrence, mais on peut également obtenir l'expression explicite de a_{k+1}:

$$a_{k+1} = \frac{(Cx_k, x_k)}{(p_k, p_k)} = \frac{(Dw_k, z_k)}{(Ds_k, s_k)}, \quad k = 0, 1, \ldots \tag{23}$$

Les formules (21), (22) et (23) décrivent le second algorithme de la méthode des directions conjuguées. Les calculs sont conduits ici dans l'ordre suivant:

1) d'après y_0 donné on calcule le résidu $r_0 = Ay_0 - f$, on résout l'équation $Bw_0 = r_0$ pour la correction w_0 et l'on pose que $s_0 = w_0$;

2) suivant la formule (23) on obtient le paramètre a_1 et on calcule $y_1 = y_0 - a_1 s_0$. Ensuite, pour $k = 1, 2, \ldots$ on exécute successivement les opérations:

3) on calcule le résidu $r_k = Ay_k - f$ et on résout l'équation de la correction $Bw_k = r_k$;

4) avec la formule (22) on calcule le paramètre b_k et on recherche s_k suivant la formule $s_k = w_k + b_k s_{k-1}$;

5) avec la formule (23) on détermine le paramètre a_{k+1} et l'approximation y_{k+1} se calcule suivant la formule

$$y_{k+1} = y_k - a_{k+1}s_k.$$

Notons que dans l'algorithme proposé il faut mémoriser y_k, w_k et s_k, c'est-à-dire le même volume de l'information intermédiaire que dans le premier algorithme.

§ 4. Exemples de méthodes à trois couches

1. Cas particuliers des méthodes des directions conjuguées. Au § 3 on a construit les méthodes itératives à trois couches des directions conjuguées utilisées pour la résolution de l'équation linéaire

$$Au = f. \tag{1}$$

Les approximations itératives se calculent suivant le schéma à trois couches

$$By_{k+1} = \alpha_{k+1} (B - \tau_{k+1}A) y_k + (1 - \alpha_{k+1}) By_{k-1} + \alpha_{k+1}\tau_{k+1}f,$$
$$k = 1, 2, \ldots, \tag{2}$$
$$By_1 = (B - \tau_1 A) y_0 + \tau_1 f, \quad y_0 \in H,$$

tandis que les paramètres d'itération α_{k+1} et τ_{k+1} s'obtiennent suivant les formules

$$\tau_{k+1} = \frac{(Dw_k, z_k)}{(Dw_k, w_k)}, \quad k = 0, 1, \ldots,$$
$$\alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(Dw_k, z_k)}{(Dw_{k-1}, z_{k-1})} \cdot \frac{1}{\alpha_k}\right)^{-1}, \tag{3}$$
$$k = 1, 2, \ldots, \quad \alpha_1 = 1,$$

où $w_k = B^{-1}r_k$ est la correction, $r_k = Ay_k - f$ le résidu, $z_k = y_k - u$ l'erreur.

Le choix des paramètres α_k et τ_k suivant les formules (3) garantit au cas d'un opérateur $DB^{-1}A$ autoadjoint et défini positif le minimum pour tout n de la norme d'erreur z_n dans H_D avec le passage de y_0 à y_n.

Examinons maintenant les cas particuliers des méthodes des directions conjuguées définis par le choix de l'opérateur D. On a vu au § 2 quatre exemples des méthodes du gradient à deux couches. A chacune de ces méthodes à deux couches correspond une méthode déterminée des directions conjuguées à trois couches. On énumérera ces méthodes en indiquant les conditions imposées aux opérateurs A et B pour obliger l'opérateur $DB^{-1}A$ à être autoadjoint. A ces méthodes s'applique le théorème 2, quant aux inégalités déterminant les constantes γ_1 et γ_2, elles seront fournies avec la description de la méthode correspondante.

1) *Méthode des gradients conjugués.*

Opérateur $D: D = A$.

Conditions: $\gamma_1 B \leqslant A \leqslant \gamma_2 B$, $\gamma_1 > 0$, $A = A^* > 0$, $B = B^* > 0$.
Formules des paramètres d'itération:

$$\tau_{k+1} = \frac{(r_k, w_k)}{(Aw_k, w_k)}, \quad \alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(r_k, w_k)}{(r_{k-1}, w_{k-1})} \frac{1}{\alpha_k}\right)^{-1}.$$

2) *Méthode des résidus conjugués.*
Opérateur D: $D = A^*A$.
Conditions: $\gamma_1 (Bx, Bx) \leqslant (Ax, Bx) \leqslant \gamma_2 (Bx, Bx)$, $\gamma_1 > 0$, $B^*A = A^*B$.

Si les hypothèses $A = A^* > 0$, $B = B^* > 0$, $AB = BA$ sont vraies, les conditions deviennent de la forme

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \quad \gamma_1 > 0.$$

Formules des paramètres d'itération:

$$\tau_{k+1} = \frac{(Aw_k, r_k)}{(Aw_k, Aw_k)}, \quad \alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(Aw_k, r_k)}{(Aw_{k-1}, r_{k-1})} \cdot \frac{1}{\alpha_k}\right)^{-1}.$$

3) *Méthode des corrections conjuguées.*
Opérateur D: $D = AB^{-1}A$.
Conditions: $\gamma_1 B \leqslant A \leqslant \gamma_2 B$, $\gamma_1 > 0$, $A = A^* > 0$, $B = B^* > 0$.
Formules des paramètres d'itération:

$$\tau_{k+1} = \frac{(Aw_k, w_k)}{(B^{-1}Aw_k, Aw_k)}, \quad \alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(Aw_k, w_k)}{(Aw_{k-1}, w_{k-1})} \cdot \frac{1}{\alpha_k}\right)^{-1}.$$

4) *Méthode des erreurs conjuguées.*
Opérateur D: $D = B_0$.
Conditions: $B = (A^*)^{-1}B_0$, $\gamma_1 B_0 \leqslant A^*A \leqslant \gamma_2 B_0$, $B_0 = B_0^* > 0$.
Formules des paramètres d'itération:

$$\tau_{k+1} = \frac{(r_k, r_k)}{(Aw_k, r_k)}, \quad \alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(r_k, r_k)}{(r_{k-1}, r_{k-1})} \cdot \frac{1}{\alpha_k}\right)^{-1}.$$

2. Méthodes à trois couches localement optimales. Revenons maintenant au procédé de construction des paramètres d'itération α_{k+1} et τ_{k+1} pour la méthode des directions conjuguées à trois couches étudié au § 3. Rappelons que les paramètres α_{k+1} et τ_{k+1} ont été choisis sur la base des conditions $(Cx_{k-1}, x_{k+1}) = 0$ et $(Cx_k, x_{k+1}) = 0$ avec l'hypothèse que les approximations itératives $y_1, y_2, \ldots, y_k$ garantissent que les conditions

$$(Cx_j, x_i) = 0, \quad j = 0, 1, \ldots, i-1, \quad i = 1, 2, \ldots, k \tag{4}$$

seront remplies.

Dans un processus de calcul idéal les conditions (4) sont remplies, aussi le choix des paramètres α_{k+1} et τ_{k+1} suivant les formules obtenues au § 3 garantit-il en fait le minimum de la norme d'erreur z_{k+1} dans H_D avec le passage de y_0 à y_{k+1}. Par contre, dans un pro-

cessus de calcul réel, qui tient compte des erreurs d'arrondi, les approximations itératives $y_1, y_2, \ldots, y_k$ seront calculées de façon imprécise et, partant, les conditions (4) ne seront pas remplies. En maintes occasions cela peut donner lieu à une diminution de la vitesse de convergence de la méthode et, quelquefois, peut même entraîner sa divergence.

Construisons à présent une modification de la méthode des directions conjuguées dénuée du défaut mentionné. En qualité de solution approchée de l'équation $Au = f$ prenons le schéma itératif à trois couches

$$By_{k+1} = \alpha_{k+1}(B - \tau_{k+1}A)\, y_k + (1 - \alpha_{k+1})\, By_{k-1} + \alpha_{k+1}\tau_{k+1}f, \quad (5)$$
$$k = 1, 2, \ldots$$

à approximations arbitraires y_0 et $y_1 \in H$. En posant y_k et y_{k-1} donnés, choisissons les paramètres α_{k+1} et τ_{k+1} sur la base de la condition du minimum de la norme d'erreur z_{k+1} dans H_D, c'est-à-dire à partir de la condition de l'optimisation locale en une seule itération du schéma à trois couches.

Ce problème sera résolu dans la seule hypothèse où l'opérateur $DB^{-1}A$ est défini positif. A cette fin passons à l'équation de l'erreur équivalente $x_k = D^{1/2}z_k$:

$$x_{k+1} = \alpha_{k+1}(E - \tau_{k+1}C)\, x_k + (1 - \alpha_{k+1})\, x_{k-1}, \quad C = D^{1/2}B^{-1}AD^{-1/2}. \quad (6)$$

Pour abréger les calculs, posons

$$1 - \alpha_{k+1} = a, \qquad \tau_{k+1}\alpha_{k+1} = b \quad (7)$$

et récrivons (6) sous la forme suivante:

$$x_{k+1} = x_k - a\,(x_k - x_{k-1}) - bCx_k. \quad (8)$$

Le problème est posé de la sorte: choisir a et b sur la base de la condition du minimum de la norme x_{k+1} dans H. Calculons la norme x_{k+1}. De (8) il s'ensuit

$$\|x_{k+1}\|^2 = \|x_k\|^2 + a^2\,\|x_k - x_{k-1}\|^2 + b^2\,\|Cx_k\|^2 - 2a\,(x_k, x_k - x_{k-1}) - 2b\,(Cx_k, x_k) + 2ab\,(Cx_k, x_k - x_{k-1}).$$

En égalant à zéro les dérivées partielles en a et b, on obtient le système relativement aux paramètres a et b

$$\begin{aligned} \|x_k - x_{k-1}\|^2 a + (Cx_k, x_k - x_{k-1})\, b &= (x_k, x_k - x_{k-1}), \\ (Cx_k, x_k - x_{k-1})\, a + \|Cx_k\|^2 b &= (Cx_k, x_k). \end{aligned} \quad (9)$$

Le déterminant du système est égal à $\|x_k - x_{k-1}\|^2\,\|Cx_k\|^2 - (Cx_k, x_k - x_{k-1})^2$ et, en vertu de l'inégalité de Cauchy-Bouniakovski, ne devient nul que quand $x_k - x_{k-1}$ est proportionnel à

$Cx_k : x_k - x_{k-1} = dCx_k$. Dans ce cas les équations du système sont proportionnelles et ce dernier se réduit à une seule équation

$$(b + ad) \| Cx_k \|^2 = (Cx_k, x_k). \tag{10}$$

Vu que dans ce cas (8) prend la forme de $x_{k+1} = x_k - (b + ad)\, Cx_k$, on obtient, en posant dans (10) $a = 0$, sur la base de (7). (10)

$$\alpha_{k+1} = 1, \quad \tau_{k+1} = \frac{(Cx_k, x_k)}{(Cx_k, Cx_k)}. \tag{11}$$

Si le déterminant n'est pas nul, alors, en résolvant le système (9), on obtient

$$a = \frac{\| Cx_k \|^2 (x_k, x_k - x_{k-1}) - (Cx_k, x_k)(Cx_k, x_k - x_{k-1})}{\| x_k - x_{k-1} \|^2 \| Cx_k \|^2 - (Cx_k, x_k - x_{k-1})^2},$$

$$b = \frac{(Cx_k, x_k)}{(Cx_k, Cx_k)}(1 - a) + \frac{(Cx_k, x_{k-1})}{(Cx_k, Cx_k)}\, a.$$

De là, en utilisant les notations (7), on aboutit aux formules des paramètres α_{k+1} et τ_{k+1} :

$$\begin{aligned} \alpha_{k+1} &= \frac{(Cx_k, x_k - x_{k-1})(Cx_k, x_{k-1}) - (x_{k-1}, x_k - x_{k-1})(Cx_k, Cx_k)}{(Cx_k, Cx_k)(x_k - x_{k-1}, x_k - x_{k-1}) - (Cx_k, x_k - x_{k-1})^2}, \\ \tau_{k+1} &= \frac{(Cx_k, x_k)}{(Cx_k, Cx_k)} + \frac{1 - \alpha_{k+1}}{\alpha_{k+1}} \frac{(Cx_k, x_{k-1})}{(Cx_k, Cx_k)}, \quad k = 1, 2, \ldots \end{aligned} \tag{12}$$

Les formules (11) obtenues auparavant peuvent être traitées comme un cas particulier des formules générales (12), en posant $\alpha_{k+1} = 1$, si le dénominateur dans l'expression de α_{k+1} devient nul.

Les formules (12) sont plus compliquées que celles des paramètres α_{k+1} et τ_{k+1} de la méthode des directions conjuguées obtenues au § 3. Dans le cas considéré, il faut de plus calculer les produits scalaires supplémentaires. Cependant le processus d'itérations (5), (12) est moins assujetti aux erreurs d'arrondi, les erreurs commises au cours des itérations précédentes s'estompent.

La liaison entre les méthodes à trois couches localement optimales et les méthodes des directions conjuguées est fixée par le

T h é o r è m e 3. *Si pour la méthode* (5), (12) *l'approximation initiale y_1 est choisie de la façon suivante*:

$$By_1 = (B - \tau_1 A) y_0 + \tau_1 f, \quad \tau_1 = \frac{(Dw_0, z_0)}{(Dw_0, w_0)}, \tag{13}$$

alors au cas où l'opérateur $DB^{-1}A$ est autoadjoint la méthode (5), (12) *coïncide avec celle des directions conjuguées.*

La démonstration s'effectuera par induction. Il s'ensuit de la condition du théorème que les approximations y_1 obtenues ici et dans la méthode des directions conjuguées coïncident. Admettons que les approximations $y_1, y_2, \ldots, y_k$ coïncident. Démontrons que y_{k+1} construit à l'aide des formules (5), (12) coïncide avec l'approximation y_{k+1} de la méthode des directions conjuguées.

D'après les hypothèses faites les paramètres d'itération $\tau_1, \tau_2, \ldots$ $\ldots, \tau_k$ et $\alpha_2, \alpha_3, \ldots, \alpha_k$ des deux méthodes coïncident également. Si l'on montre que les paramètres τ_{k+1} et α_{k+1} de ces méthodes coïncident de même, la proposition du théorème 3 sera démontrée.

Vu que $y_1, y_2, \ldots, y_k$ sont des approximations itératives de la méthode des directions conjuguées, en vertu du lemme les conditions

$$(Cx_j, x_i) = 0, \quad j = 0, 1, \ldots, i-1, \quad i = 1, 2, \ldots, k \tag{14}$$

sont remplies. En portant (14) avec $j = k - 1$ et $i = k$ dans (12) et utilisant l'opérateur autoadjoint C, il vient

$$\alpha_{k+1} = \frac{(x_{k-1}, x_k - x_{k-1})(Cx_k, Cx_k)}{(Cx_k, x_k)^2 - \| Cx_k \|^2 \| x_k - x_{k-1} \|^2}, \quad \tau_{k+1} = \frac{(Cx_k, x_k)}{(Cx_k, Cx_k)}. \tag{15}$$

Ainsi donc les paramètres τ_{k+1} de la méthode localement optimale et de celle des directions conjuguées coïncident. Il ne reste qu'à montrer que les paramètres α_{k+1} coïncident aussi.

De (6) et (13) il vient

$$\begin{aligned} x_k - x_{k-1} &= (\alpha_k - 1)(x_{k-1} - x_{k-2}) - \alpha_k \tau_k C x_{k-1}, \\ & \qquad\qquad k = 2, 3, \ldots, \\ x_1 - x_0 &= -\tau_1 C x_0. \end{aligned} \tag{16}$$

De (16) il s'ensuit que la différence $x_k - x_{k-1}$ est une combinaison linéaire $Cx_0, Cx_1, \ldots, Cx_{k-1}$ qui prend la forme suivante

$$\begin{aligned} x_k - x_{k-1} &= -\alpha_k \tau_k C x_{k-1} + \sum_{j=0}^{k-2} \beta_j C x_j, \quad k \geqslant 2, \\ x_1 - x_0 &= -\tau_1 C x_0, \end{aligned} \tag{17}$$

où les coefficients β_j s'expriment en fonction de $\tau_1, \tau_2, \ldots, \tau_{k-1}$ et $\alpha_2, \alpha_3, \ldots, \alpha_{k-1}$. En multipliant le premier et le second membre de (17) scalairement par x_{k-1} et $x_k - x_{k-1}$ et compte tenu de (14), il vient

$$\begin{aligned} (x_{k-1}, x_k - x_{k-1}) &= -\alpha_k \tau_k (Cx_{k-1}, x_{k-1}), \\ \| x_k - x_{k-1} \|^2 &= \alpha_k \tau_k (Cx_{k-1}, x_{k-1}). \end{aligned} \tag{18}$$

En portant (18) dans l'expression (15) pour α_{k+1} et compte tenu du paramètre τ_{k+1}, on obtient la formule

$$\alpha_{k+1} = \frac{\alpha_k \tau_k (Cx_{k-1}, x_{k-1})(Cx_k, Cx_k)}{\alpha_k \tau_k (Cx_{k-1}, x_{k-1})(Cx_k, Cx_k) - (Cx_k, x_k)^2} = \left(1 - \frac{\tau_{k+1}(Cx_k, x_k)}{\tau_k (Cx_{k-1}, x_{k-1})} \cdot \frac{1}{\alpha_k}\right)^{-1}$$

coïncidant avec la formule obtenue pour le paramètre α_{k+1} dans la méthode des directions conjuguées. Le théorème est démontré.

En portant $x_k = D^{1/2}z_k$ et $C = D^{-1/2}(DB^{-1}A)D^{-1/2}$ dans (12), on obtient la forme suivante des formules pour les paramètres α_{k+1} et τ_{k+1}:

$$\begin{aligned}\alpha_{k+1} &= \frac{(Dw_k, z_k - z_{k-1})(Dw_k, z_{k-1}) - (Dz_{k-1}, y_k - y_{k-1})(Dw_k, w_k)}{(Dw_k, w_k)(D(z_k - z_{k-1}), y_k - y_{k-1}) - (Dw_k, z_k - z_{k-1})^2}, \\ \tau_{k+1} &= \frac{(Dw_k, z_k)}{(Dw_k, w_k)} + \frac{1-\alpha_{k+1}}{\alpha_{k+1}} \frac{(Dw_k, z_{k-1})}{(Dw_k, w_k)}.\end{aligned} \tag{19}$$

Si l'on introduit les notations pour les produits scalaires

$$a_k = (Dw_k, z_k), \quad b_k = (Dw_k, z_{k-1}), \quad c_k = (Dz_k, y_k - y_{k-1}),$$
$$d_k = (Dz_{k-1}, y_k - y_{k-1}), \quad e_k = (Dw_k, w_k),$$

les formules (19) se récriront sous la forme

$$\alpha_{k+1} = \frac{(a_k - b_k)b_k - d_k e_k}{(c_k - d_k)e_k - (a_k - b_k)^2}, \quad k = 1, 2, \ldots, \quad \alpha_1 = 1,$$
$$\tau_{k+1} = \frac{a_k}{e_k} + \frac{1-\alpha_{k+1}}{\alpha_{k+1}} \frac{b_k}{e_k}, \quad k = 0, 1, \ldots$$

Fournissons les expressions de a_k, b_k, c_k, d_k et e_k pour des choix concrets de l'opérateur D:

1) $D = A$, $A = A^*$

$$a_k = (w_k, r_k), \quad b_k = (w_k, r_{k-1}), \quad c_k = (r_k, y_k - y_{k-1}),$$
$$d_k = (r_{k-1}, y_k - y_{k-1}), \quad e_k = (Aw_k, w_k);$$

2) $D = A^*A$

$$a_k = (Aw_k, r_k), \quad b_k = (Aw_k, r_{k-1}), \quad c_k = (r_k, r_k - r_{k-1}),$$
$$d_k = (r_{k-1}, r_k - r_{k-1}), \quad e_k = (Aw_k, Aw_k);$$

3) $D = A^*B^{-1}A$, $B = B^*$

$$a_k = (Aw_k, w_k), \quad b_k = (Aw_k, w_{k-1}), \quad c_k = (w_k, r_k - r_{k-1}),$$
$$d_k = (w_{k-1}, r_k - r_{k-1}), \quad e_k = (B^{-1}Aw_k, Aw_k).$$

§ 5. Accélération de la convergence des méthodes à deux couches au cas d'un opérateur autoadjoint

1. Algorithme du processus d'accélération. Au point 5 du § 1 il a été établi qu'au cas d'un opérateur autoadjoint $DB^{-1}A$ les méthodes du gradient à deux couches possèdent la propriété asymptotique. Cette dernière se manifeste dans le fait que pour des grands numéros d'itérations la vitesse de convergence de la méthode diminue sensiblement vis-à-vis du début d'itérations. On a également montré que pour des grands numéros d'itérations les erreurs observées toutes les deux itérations deviennent presque proportionnelles.

En utilisant cette propriété, passons maintenant à la construction du processus d'accélération de la convergence des méthodes du gradient à deux couches.

Pour la résolution de l'équation

$$Au = f \tag{1}$$

prenons la méthode itérative du gradient à deux couches

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}}+Ay_k=f, \quad k=0, 1, \ldots, \quad y_0 \in H, \tag{2}$$

$$\tau_{k+1}=\frac{(Dw_k, z_k)}{(Dw_k, w_k)}, \quad k=0, 1, \ldots \tag{3}$$

Posons que l'opérateur $DB^{-1}A$ est autoadjoint dans H. La méthode itérative possède alors la propriété asymptotique et, pour un numéro d'itérations k suffisamment grand, on a l'égalité approchée

$$z_{k+2} \approx \rho^2 z_k, \quad z_k = y_k - u. \tag{4}$$

Voyons d'abord le cas où dans (4) on a une égalité stricte, c'est-à-dire que $z_{k+2} = \rho^2 z_k$. Construisons sur la base des approximations déjà trouvées y_k et y_{k+2} la nouvelle approximation suivant la formule

$$y = \alpha y_{k+2} + (1 - \alpha)y_k, \quad \alpha = 1/(1 - \rho^2). \tag{5}$$

On obtient pour l'erreur $z = y - u$

$$z = \alpha z_{k+2} + (1 - \alpha)\, z_k = (\alpha\rho^2 + 1 - \alpha)\, z_k = \\ = [1 - \alpha\,(1 - \rho^2)]z_k = 0.$$

Par conséquent, en cas de la réalisation de l'égalité stricte (4), la combinaison linéaire (5) des approximations y_k et y_{k+2} fournit une solution précise de l'équation (1).

Comme il a été noté au § 1, la réalisation de l'égalité stricte (4) constitue un cas exceptionnel qui n'a lieu que pour une approximation initiale spéciale. Dans le cas général on a l'égalité approchée (4), tandis que les arguments avancés plus haut permettent d'espérer qu'une certaine combinaison linéaire de y_k et y_{k+2} fournira une bonne approximation à la solution du problème initial.

Cherchons la meilleure entre ces combinaisons linéaires. Soient y_k, y_{k+1} et y_{k+2} les approximations itératives obtenues suivant les formules (2), (3). Cherchons la nouvelle approximation y suivant la formule

$$y = \alpha y_{k+2} + (1 - \alpha)\, y_k. \tag{6}$$

Posons le problème où le paramètre α est choisi de façon que la norme d'erreur $z = y - u$ dans H_D soit minimale.

En utilisant d'abord le schéma (2), éliminons y_{k+2} de (6). Il vient

$$By_{k+2} = (B - \tau_{k+2}A)y_{k+1} + \tau_{k+2}f$$

et. après avoir porté y_{k+2} dans (6). on obtient

$$By = \alpha (B - \tau_{k+2}A) y_{k+1} + (1 - \alpha) By_k + \alpha\tau_{k+2} f, \qquad (7)$$

où y_{k+1} s'obtient suivant le schéma à deux couches

$$By_{k+1} = (B - \tau_{k+1}A) y_k + \tau_{k+1}f. \qquad (8)$$

Si l'on admet que y_k est l'approximation initiale donnée, le schéma (7), (8) coïncide alors avec le schéma de la méthode des directions conjuguées, les paramètres τ_{k+1} et τ_{k+2} étant les mêmes que ceux de la méthode des directions conjuguées. Il s'ensuit de la théorie de cette méthode (voir point 1, § 4, formule (3)) que la valeur optimale du paramètre α se détermine par la formule

$$\alpha = \frac{1}{1 - \frac{\tau_{k+2}}{\tau_{k+1}} \frac{(Dw_{k+1}, z_{k+1})}{(Dw_k, z_k)}}. \qquad (9)$$

Bref, le problème posé du meilleur choix du paramètre α est ainsi résolu. Les formules (6), (9) définissent le procédé d'accélération.

Notons que pour déterminer y il est possible, au lieu d'utiliser la formule (6), de le calculer à l'aide du schéma à deux couches suivant:

$$\begin{aligned} B\bar{y}_{k+1} &= (B - \bar{\tau}_{k+1}A) y_k + \bar{\tau}_{k+1}f, \\ By &= (B - \bar{\tau}_{k+2}A) \bar{y}_{k+1} + \bar{\tau}_{k+2}f, \end{aligned} \qquad (10)$$

où $\bar{\tau}_{k+1}$ et $\bar{\tau}_{k+2}$ sont les racines de l'équation

$$\tau^2 - \alpha (\tau_{k+1} + \tau_{k+2}) \tau + \alpha\tau_{k+1}\tau_{k+2} = 0.$$

Pour $\bar{\tau}_{k+1}$ il faut choisir la racine minimale.

L'utilisation de (10) au lieu de (6) dispense d'augmenter le volume de l'information intermédiaire mémorisée.

2. Appréciation de l'efficience. Apprécions maintenant l'efficience du procédé d'accélération. Avant de calculer la norme d'erreur $z = y - u$ dans H_D, transformons l'expression (9) pour α.

La substitution $z_k = D^{-1/2}x_k$ dans (9) donne

$$\alpha = \left(1 - \frac{\tau_{k+2}}{\tau_{k+1}} \frac{(Cx_{k+1}, x_{k+1})}{(Cx_k, x_k)}\right)^{-1}, \quad C = D^{1/2}B^{-1}AD^{-1/2}. \qquad (11)$$

De (10) et (11), § 1, il vient

$$\| x_{k+1} \| = \rho_{k+1} \| x_k \|, \quad \rho_{k+1}^2 = 1 - \frac{(Cx_k, x_k)^2}{(Cx_k, Cx_k) \| x_k \|^2}. \qquad (12)$$

De la formule (9), § 1, il vient

$$\tau_{k+1} = \frac{(Cx_k, x_k)}{(Cx_k, Cx_k)}. \qquad (13)$$

Utilisant (12) et (13), il vient

$$\frac{\tau_{k+2}}{\tau_{k+1}} \cdot \frac{(Cx_{k+1}, x_{k+1})}{(Cx_k, x_k)} = \frac{1-\rho_{k+2}^2}{1-\rho_{k+1}^2} \rho_{k+1}^2.$$

En portant cette expression dans (11), on obtient

$$\alpha = \frac{1-\rho_{k+1}^2}{1-2\rho_{k+1}^2+\rho_{k+1}^2\rho_{k+2}^2}. \qquad (14)$$

Calculons maintenant la norme d'erreur $z = y - u$ dans H_D. De (6) il vient

$$z = \alpha z_{k+2} + (1-\alpha) z_k.$$

De là, pour l'erreur équivalente $z_k = D^{1/2}x_k$ et $x = D^{1/2}z$, on aura $x = \alpha x_{k+2} + (1-\alpha) x_k$. Calculons la norme x dans H. Il vient

$$\|x\|^2 = \alpha^2 \|x_{k+2}\|^2 + 2\alpha(1-\alpha)(x_{k+2}, x_k) + (1-\alpha)^2 \|x_k\|^2.$$

De l'égalité $(x_{k+2}, x_k) = \|x_{k+1}\|^2$ démontrée au point 5, § 1, on tire

$$\|x\|^2 = \alpha^2 \|x_{k+2}\|^2 + 2\alpha(1-\alpha)\|x_{k+1}\|^2 + (1-\alpha)^2 \|x_k\|^2.$$

En y portant l'expression (14) pour α et utilisant (12), il vient

$$\|x\|^2 = \frac{\rho_{k+2}^2-\rho_{k+1}^2}{\rho_{k+1}^2(1-2\rho_{k+1}^2+\rho_{k+1}^2\rho_{k+2}^2)} \|x_{k+2}\|^2 < \|x_{k+2}\|^2. \qquad (15)$$

Vu que $\rho_{k+1} \leqslant \rho_{k+2} \leqslant \rho < 1$, on a

$$1-2\rho_{k+1}^2+\rho_{k+1}^2\rho_{k+2}^2 \geqslant (1-\rho^2)^2,$$

par conséquent, on a pour la norme x l'estimation

$$\|x\|^2 \leqslant \left(\frac{\rho_{k+2}^2}{\rho_{k+1}^2}-1\right) \frac{\|x_{k+2}\|^2}{(1-\rho^2)^2}.$$

En vertu de la propriété asymptotique pour des numéros de k suffisamment grands, on a $\rho_{k+1} \approx \rho_{k+2}$, l'effet du procédé d'accélération devenant sensible.

Notons que bien que l'accélération efficace de la convergence se manifeste pour des grands numéros d'itérations k, ce procédé peut être également utilisé pour tout numéro d'itérations. Il est recommandé d'arrêter de temps en temps le processus d'itérations mené suivant le schéma à deux couches (2), (3) et de calculer la nouvelle approximation par le procédé proposé. Le processus d'itérations peut être arrêté avec le calcul d'une telle approximation si pour y_{k+2} obtenu l'inégalité

$$\frac{\rho_{k+2}^2-\rho_{k+1}^2}{\rho_{k+1}^2(1-2\rho_{k+1}^2+\rho_{k+1}^2\rho_{k+2}^2)} \|z_{k+2}\|_D^2 \leqslant \varepsilon^2 \|z_0\|_D^2$$

sera vérifiée. En effet, dans ce cas, en vertu de (15), on obtient $\|y-u\|_D \leqslant \varepsilon \|y_0-u\|_D$, c'est-à-dire que la précision exigée ε sera atteinte.

3. Exemple. Afin d'illustrer l'efficience du procédé proposé d'accélération de la convergence des méthodes du gradient à deux couches, examinons la solution du problème modèle par la méthode implicite de la plus grande pente. En guise d'exemple, prenons le problème discret de Dirichlet pour l'équation de Laplace sur maillage carré $\bar{\omega} = \{x_{ij} = (ih, jh),\ 0 \leqslant i \leqslant N,\ 0 \leqslant j \leqslant N,\ h = 1/N\}$ dans un carré unitaire

$$\begin{gathered} \Lambda u = \Lambda_1 u + \Lambda_2 u = 0, \quad x \in \omega, \quad u|_\gamma = 0, \\ \Lambda_\alpha u = u_{\bar{x}_\alpha x_\alpha}, \quad \alpha = 1, 2. \end{gathered} \tag{16}$$

Introduisons l'espace H, composé de fonctions de mailles données sur ω avec produit scalaire

$$(u, v) = \sum_{x \in \omega} u(x)\, v(x)\, h^2.$$

L'opérateur A se définit de la façon suivante: $A = A_1 + A_2$, $A_\alpha y = -\Lambda_\alpha v$, $y \in H$, où $v(x) = y(x)$ pour $x \in \omega$ et $v|_\gamma = 0$.

Ecrivons le problème (16) sous forme d'équation opératorielle

$$Au = f, \quad f = 0. \tag{17}$$

En guise d'opérateur B choisissons l'opérateur factorisé suivant: $B = (E + \omega A_1)(E + \omega A_2)$, $\omega > 0$, où ω est le paramètre d'itération.

Les opérateurs A_1 et A_2 étant autoadjoints et permutables dans H, les opérateurs A et B sont autoadjoints dans H. En outre, on montre sans peine que les opérateurs A et B sont définis positifs dans H. Par conséquent, pour résoudre l'équation (17) il est possible d'utiliser la méthode implicite de la plus grande pente

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad \tau_{k+1} = \frac{(w_k, r_k)}{(Aw_k, w_k)}, \quad k = 0, 1, \ldots \tag{18}$$

Dans cette méthode $D = A$ et $DB^{-1}A = AB^{-1}A$. L'opérateur $DB^{-1}A$ étant autoadjoint dans H, la méthode étudiée possède la propriété asymptotique. Il s'ensuit de la théorie de la méthode de la plus grande pente (voir point 1, § 2) que la vitesse de convergence de la méthode se définit dans ce cas par la relation $\xi = \gamma_1/\gamma_2$, où γ_1 et γ_2 sont des constantes de l'équivalence énergétique des opérateurs A et B: $\gamma_1 B \leqslant A \leqslant \gamma_2 B$, $\gamma_1 > 0$.

Le paramètre d'itération ω est donc choisi sur la base de la condition du maximum de ξ. Dans le § 2 du chapitre XI on montrera que la valeur optimale de ω est déterminée suivant la formule

$$\omega = \frac{1}{\sqrt{\delta\Delta}}, \quad \delta = \frac{4}{h^2} \sin^2 \frac{\pi h}{2}, \quad \Delta = \frac{4}{h^2} \cos^2 \frac{\pi h}{2},$$

de plus,

$$\gamma_1 = \frac{2\delta}{(1 + \sqrt{\eta})^2}, \quad \gamma_2 = \frac{(\Delta + \delta)\sqrt{\eta}}{(1 + \sqrt{\eta})^2}, \quad \xi = \frac{2\sqrt{\eta}}{1 + \eta}, \quad \eta = \frac{\delta}{\Delta}.$$

Pour l'exemple étudié

$$\omega=\frac{h^2}{2\sin\pi h},\quad \gamma_1=\frac{2}{h^2}\,\frac{\sin\pi h}{1+\sin\pi h},\quad \gamma_2=\frac{2}{h^2}\,\frac{\sin\pi h}{1+\sin\pi h},\quad \xi=\sin\pi h.$$

Donnons les résultats des calculs au cas où l'approximation initiale y_0 est choisie égale à $y_0(x) = e^{(x_1-x_2)}$ pour $x \in \omega$, $y_0|_\gamma = 0$. La précision exigée ε était prise égale à 10^{-4}, $N = 40$.

Au tableau 8 pour quelques numéros d'itérations k sont fournis: $\| z_k \|_D / \| z_0 \|_D$ qui est la précision relative de la k-ième itération; $\rho_k = \| z_k \|_D / \| z_{k-1} \|_D$, grandeur caractérisant la diminution de la norme d'erreur avec le passage de la $(k-1)$-ième itération à la k-ième; $\gamma_1^{(k)}$ et $\gamma_2^{(k)}$. valeurs approchées de γ_1 et γ_2 obtenues comme des racines de l'équation du second degré

$$(1-\tau_k\gamma)(1-\tau_{k-1}\gamma)=\rho_k\rho_{k-1},\quad k=2, 3, \ldots,$$

et les paramètres d'itération τ_k.

Tableau 8

k	$\|z_k\|_D/\|z_0\|_D$	ρ_k	$\gamma_1^{(k)}$	$\gamma_2^{(k)}$	τ_k
1	$3{,}6\cdot 10^{-1}$	0,36203	—	—	$5{,}392\cdot 10^{-3}$
2	$2{,}3\cdot 10^{-1}$	0,63810	77,31858	236,1883	$7{,}809\cdot 10^{-3}$
3	$1{,}8\cdot 10^{-1}$	0,76998	40,59796	232,1435	$6{,}911\cdot 10^{-3}$
4	$1{,}4\cdot 10^{-1}$	0,81178	26,87824	233,4976	$8{,}644\cdot 10^{-3}$
. . .	. . .	. . .	. . .	. . .	. . .
26	$3{,}9\cdot 10^{-3}$	0,85175	18,27141	230,5962	$8{,}876\cdot 10^{-3}$
27	$3{,}4\cdot 10^{-3}$	0,85178	18,26983	230,6607	$7{,}338\cdot 10^{-3}$
28	$2{,}9\cdot 10^{-3}$	0,85183	18,27026	230,7191	$8{,}872\cdot 10^{-3}$
29	$2{,}4\cdot 10^{-3}$	0,85186	18,26895	230,7771	$7{,}335\cdot 10^{-3}$
. . .	. . .	. . .	. . .	. . .	. . .
46	$1{,}6\cdot 10^{-4}$	0,85226	18,26677	231,4121	$8{,}845\cdot 10^{-3}$
47	$1{,}4\cdot 10^{-4}$	0,85227	18,26632	231,4375	$7{,}318\cdot 10^{-3}$
48	$1{,}2\cdot 10^{-4}$	0,85229	18,26664	231,4612	$8{,}843\cdot 10^{-3}$
49	$9{,}9\cdot 10^{-4}$	0,85230	18,26623	231,4849	$7{,}317\cdot 10^{-3}$

La précision exigée ε a été atteinte après 49 itérations suivant le schéma (18). Pour $\varepsilon = 10^{-4}$ le nombre théorique d'itérations est 59. Les valeurs de ρ_k données au tableau illustrent de façon satisfaisante la propriété asymptotique de la méthode. On voit qu'avec l'accroissement du numéro d'itérations la vitesse de convergence de la méthode diminue. La précision de $4\cdot 10^{-3}$ a été atteinte en 26 itérations, mais l'accroissement de la précision de 40 fois a exigé 23 itérations supplémentaires. La quantité ρ_k croît de façon monotone et pour $k = 26$ on a $\rho_{k+1} - \rho_k \approx 3\cdot 10^{-5}$. Les paramètres d'itération τ_k et τ_{k+2} deviennent presque égaux.

Afin de comparer les valeurs approchées $\gamma_1^{(k)}$ et $\gamma_2^{(k)}$ avec les valeurs précises, donnons celles de γ_1 et γ_2:

$$\gamma_1 = 18{,}26556, \quad \gamma_2 = 232{,}8036.$$

Après 49 itérations on a trouvé γ_1 à la précision de 0,004 % près et γ_2 à la précision de 0,6 % près.

Pour accélérer la convergence de la méthode, on a eu recourt au procédé d'accélération décrit au point 1. Sur la base des approximations y_{26} et y_{28} obtenues suivant le schéma (18) on a construit, à l'aide des formules (6), (9), la nouvelle approximation y. La précision exigée $\varepsilon = 10^{-4}$ a été atteinte. L'utilisation du procédé d'accélération de la convergence des méthodes du gradient à deux couches construit dans ce paragraphe a permis de diminuer le nombre d'itérations exigées pour l'exemple analysé de près de 1,8 fois.

CHAPITRE IX

MÉTHODES ITÉRATIVES TRIANGULAIRES

On étudie dans ce chapitre les méthodes itératives à deux couches implicites à l'opérateur B desquelles correspondent des matrices triangulaires. Au § 1 on analyse la méthode de Seidel pour laquelle on formule les conditions suffisantes de convergence. Au § 2 est étudiée la méthode de surrelaxation. On y fournit le mode de choix du paramètre d'itération et on y obtient l'estimation du rayon spectral de l'opérateur de passage. Au § 3 on examine le schéma itératif général des méthodes triangulaires, on indique le choix du paramètre d'itération et on démontre la convergence de la méthode dans la norme H_A.

§ 1. Méthode de Seidel

1. Schéma itératif de la méthode. Dans les chapitres précédents on a exposé la théorie générale des méthodes itératives à deux et à trois couches utilisées pour la recherche de la solution approchée de l'équation linéaire opératorielle de première espèce

$$Au = f. \tag{1}$$

Cette théorie indique comment choisir les paramètres d'itération et fournit l'estimation du nombre d'itérations des méthodes correspondantes, de plus, cette théorie recourt à un minimum d'information de nature générale sur les opérateurs du schéma itératif. En refusant de se référer à la structure concrète des opérateurs du schéma itératif, on est en mesure de développer la théorie dans une optique unique et de construire des méthodes itératives implicites, optimales sur la classe des opérateurs B.

Dans la théorie générale des méthodes itératives on a montré que l'efficience de la méthode dépend de façon essentielle du choix de l'opérateur B. Du choix de cet opérateur dépendent aussi bien le nombre d'itérations qu'il est nécessaire d'accomplir pour atteindre la précision exigée ε que celui d'opérations arithmétiques dépensées à la mise en œuvre d'une seule itération. Chacune de ces grandeurs séparément ne peut servir de critère d'efficience de la méthode itérative. Eclairons cette assertion. Soient A et B des opérateurs autoadjoints et définis positifs dans H. Il s'ensuit de la théorie des méthodes itératives que si, en guise d'opérateur D, on prend l'un des opérateurs A, B ou $AB^{-1}A$, le nombre d'itérations des méthodes itéra-

tives passées en revue dans les chapitres VI-VIII (méthode de Tchébychev, méthode itérative simple, méthodes du type variationnel, etc.) sera alors défini par la relation $\xi = \gamma_1/\gamma_2$, où γ_1 et γ_2 sont des constantes de l'équivalence énergétique des opérateurs A et B: $\gamma_1 B \leqslant A \leqslant \gamma_2 B$.

Aussi si l'on pose $B = A$, obtient-on la valeur maximale possible $\xi = 1$ et les méthodes itératives donneront-elles une solution précise de l'équation (1) après une seule itération pour toute approximation initiale. Par conséquent, le choix fait de l'opérateur B minimise le nombre d'itérations. Cependant, pour la mise en œuvre de cette unique opération d'itération il faut inverser B, c'est-à-dire l'opérateur A. Dans ce cas, évidemment, le nombre d'opérations arithmétiques sera maximal.

D'un autre côté, pour des schémas explicites, avec $B = E$, le nombre d'opérations arithmétiques exigées par itération est minimal, mais le nombre de ces itérations s'avère alors trop grand.

Bref, il se pose un problème de choix optimal de l'opérateur B sur la base de la minimisation du volume total des calculs nécessaires à l'obtention de la solution avec la précision fixée.

Il va de soi qu'une telle position générale ne permet pas de résoudre ce problème. Actuellement le développement des méthodes itératives est orienté vers la construction d'opérateurs B facilement invertibles parmi lesquels on choisit ceux dont le rapport γ_1/γ_2 est le meilleur. Les opérateurs facilement invertibles ou économiques sont les opérateurs dont l'inversion peut être réalisée en un nombre d'opérations arithmétiques proportionnel ou presque proportionnel au nombre d'inconnues. Des exemples de tels opérateurs nous sont fournis par les opérateurs auxquels correspondent des matrices diagonale, tridiagonale, triangulaires, ainsi que leurs produits. En guise d'exemple plus compliqué, citons l'opérateur de différences de Laplace dans un rectangle qui, comme on l'a montré au chapitre IV, peut être inverti par des méthodes directes avec de petites dépenses en opérations arithmétiques.

Il faut noter que l'utilisation des opérateurs diagonaux en guise d'opérateur B permet de réduire le nombre d'itérations par rapport au cas du schéma itératif explicite. Mais l'ordre asymptotique de dépendance du nombre d'itérations de celui d'inconnues demeure le même que pour le schéma explicite. Plus alléchante est la perspective d'utilisation des opérateurs B triangulaires.

Dans le présent chapitre ainsi qu'au chapitre X on étudiera les méthodes itératives implicites universelles à deux couches, où à l'opérateur B sont associés des matrices triangulaires (méthodes triangulaires) ou le produit de matrices triangulaires (méthode triangulaire alternée).

L'étude de ces méthodes sera abordée par la méthode la plus simple, la *méthode de Seidel.*

Examinons le système d'équations algébriques linéaires (1) ou sous forme développée

$$\sum_{j=1}^{M} a_{ij}u_j = f_i, \quad i = 1, 2, \ldots, M.$$

Dans le cas considéré on a affaire à l'équation (1) donnée dans l'espace de dimension finie $H = H_M$.

La méthode itérative de Seidel dans l'hypothèse que les éléments diagonaux de la matrice $A = (a_{ij})$ sont différents de zéro $(a_{ii} \neq 0)$ s'écrit de la façon suivante:

$$\sum_{j=1}^{i} a_{ij}y_j^{(k+1)} + \sum_{j=i+1}^{M} a_{ij}y_j^{(k)} = f_i, \quad i = 1, 2, \ldots, M, \tag{2}$$

où $y_j^{(k)}$ est la j-ième composante de l'approximation itérative du numéro k. En guise d'approximation initiale on choisit un vecteur quelconque.

La définition de la $(k + 1)$-ième itération commence avec $i = 1$:

$$a_{11}y_1^{(k+1)} = -\sum_{j=2}^{M} a_{1j}y_j^{(k)} + f_1.$$

Comme $a_{11} \neq 0$, on en tire $y_1^{(k+1)}$. Pour $i = 2$, on obtient

$$a_{22}y_2^{(k+1)} = -a_{21}y_1^{(k+1)} - \sum_{j=3}^{M} a_{2j}y_j^{(k)} + f_2.$$

Posons que $y_1^{(k+1)}, y_2^{(k+1)}, \ldots, y_{i-1}^{(k+1)}$ sont déjà trouvés. Alors $y_i^{(k+1)}$ s'obtient de l'équation

$$a_{ii}y_i^{(k+1)} = -\sum_{j=1}^{i-1} a_{ij}y_j^{(k+1)} - \sum_{j=i+1}^{M} a_{ij}y_j^{(k)} + f_i. \tag{3}$$

D'après la formule (3) on voit que l'algorithme de la méthode de Seidel est particulièrement simple. La valeur de $y_i^{(k+1)}$, obtenue suivant la formule (3), se place à l'endroit de $y_i^{(k)}$.

Apprécions le nombre d'opérations arithmétiques que vaut la mise en œuvre d'une seule itération. Si tous les a_{ij} ne sont pas nuls, les calculs suivant la formule (3) exigent $M - 1$ additions, $M - 1$ multiplications et une division. Donc la mise en œuvre d'une itération vaut $2M^2 - M$ opérations arithmétiques.

Si sur chaque ligne de la matrice A seuls m éléments sont différents de zéro, et c'est justement la situation observée pour les équations de mailles elliptiques, alors la mise en œuvre d'une itération exigera $2mM - M$ opérations, autrement dit le nombre d'opérations proportionnel à celui d'inconnues M.

Ecrivons maintenant la méthode itérative de Seidel (2) sous une forme matricielle. Pour cela, représentons la matrice A sous forme d'une somme de matrices diagonale, triangulaires inférieure et supérieure

$$A = \mathcal{D} + L + U, \tag{4}$$

où

$$L = \begin{Vmatrix} 0 & & & & \\ a_{21} & 0 & & & 0 \\ a_{31} & a_{32} & 0 & & \\ \cdot & \cdot & \cdot & & \\ \cdot & & & \cdot & \\ \cdot & & & \cdot & \cdot \\ a_{M1} & a_{M2} & \dots & a_{MM-1} & 0 \end{Vmatrix}, \quad U = \begin{Vmatrix} 0 & a_{12} & a_{13} & \dots & a_{1M} \\ & 0 & a_{23} & \dots & a_{2M} \\ & & \cdot & & \cdot \\ & & & \cdot & \cdot \\ & & & \cdot & \cdot \\ & & & 0 & a_{M-1M} \\ & 0 & & & 0 \end{Vmatrix}.$$

$$\mathcal{D} = \begin{Vmatrix} a_{11} & & & 0 \\ & a_{22} & & \\ & & \ddots & \\ 0 & & & a_{MM} \end{Vmatrix}.$$

Désignons par $y_k = (y_1^{(k)}, y_2^{(k)}, \dots, y_M^{(k)})$ le vecteur de la k-ième approximation itérative.

Utilisant ces notations, écrivons la méthode de Seidel sous la forme

$$(\mathcal{D} + L)\, y_{k+1} + U y_k = f, \quad k = 0, 1, \dots$$

Réduisons ce schéma itératif à la forme canonique des schémas à deux couches

$$(\mathcal{D} + L)\,(y_{k+1} - y_k) + A y_k = f, \quad k = 0, 1, \dots, \quad y_0 \in H. \tag{5}$$

En confrontant (5) à la forme canonique

$$B\,\frac{y_{k+1} - y_k}{\tau_{k+1}} + A y_k = f, \quad k = 0, 1, \dots, \quad y_0 \in H,$$

on obtient que $B = \mathcal{D} + L$, $\tau_k \equiv 1$. Le schéma (5) est implicite, l'opérateur B est une matrice triangulaire et, partant, n'est pas autoadjoint dans H.

On a examiné la méthode de Seidel dite ponctuelle ou scalaire, en posant que les éléments a_{ij} de la matrice A sont des nombres. De façon analogue on construit la méthode de Seidel par blocs ou vectorielle pour le cas où a_{ii} sont des matrices carrées, en général, de dimension variée, tandis que a_{ij} pour $i \neq j$ sont des matrices rectangulaires. Dans ce cas y_i et f_i sont des vecteurs dont la dimension est celle de la matrice a_{ii}.

Dans l'hypothèse des matrices a_{ii} non dégénérées la méthode de Seidel par blocs s'écrit sous forme (2) ou sous forme canonique (5).

2. Exemples d'application de la méthode. Examinons comment est utilisée la méthode de Seidel à l'obtention de la solution approchée du problème discret de Dirichlet pour l'équation de Poisson et de l'équation elliptique à coefficients variables dans un rectangle.

Soit sur un maillage rectangle $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G},\ 0 \leqslant i \leqslant N_1,\ 0 \leqslant j \leqslant N_2,\ h_\alpha = l_\alpha/N_\alpha,\ \alpha = 1, 2\}$, introduit dans un rectangle $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha,\ \alpha = 1, 2\}$, il s'agit de rechercher la solution du problème discret de Dirichlet pour l'équation de Poisson

$$\Lambda y = \sum_{\alpha=1}^{2} y_{\bar{x}_\alpha x_\alpha} = -\varphi(x), \quad x \in \omega, \quad y(x) = g(x), \quad x \in \gamma, \qquad (6)$$

où $\gamma = \{x_{ij} \in \Gamma\}$ est la frontière du maillage $\bar{\omega}$.

Dans l'exemple donné les inconnues sont $y(i, j) = y(x_{ij})$ aux nœuds internes du maillage. Si l'on ordonne les inconnues de façon naturelle suivant les lignes du maillage ω, en commençant par la ligne du bas, le schéma aux différences (6) pourra alors s'écrire sous forme du système suivant d'équations algébriques:

$$-\frac{1}{h_1^2} y(i-1, j) - \frac{1}{h_2^2} y(i, j-1) + \left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right) y(i, j) -$$
$$- \frac{1}{h_1^2} y(i+1, j) - \frac{1}{h_2^2} y(i, j+1) = \varphi(i, j)$$

pour $i = 1, 2, \ldots, N_1 - 1$, $j = 1, 2, \ldots, N_2 - 1$ et $y(x) = g(x)$ avec $x \in \gamma$. Les inconnues $y(i-1, j)$ et $y(i, j-1)$ précèdent $y(i, j)$, tandis que $y(i+1, j)$ et $y(i, j+1)$ suivent $y(i, j)$. Comme dans chaque équation sont liées cinq inconnues au maximum, sur chaque ligne de la matrice A il n'y a pas plus de cinq éléments non nuls.

Pour le système envisagé la méthode de Seidel ponctuelle prendra la forme

$$\left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right) y_{k+1}(i, j) = \frac{1}{h_1^2} y_{k+1}(i-1, j) + \frac{1}{h_2^2} y_{k+1}(i, j-1) +$$
$$+ \frac{1}{h_1^2} y_k(i+1, j) + \frac{1}{h_2^2} y_k(i, j+1) + \varphi(i, j),$$
$$1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant j \leqslant N_2 - 1,$$

avec $y_k(x) = g(x)$, $x \in \gamma$ pour tout $k \geqslant 0$.

Les calculs commencent avec le point $i = 1$, $j = 1$ et se poursuivent soit par lignes, soit par colonnes du maillage ω. Pour la recherche de $y_{k+1}(i, j)$ il faut 7 opérations arithmétiques, et en tout pour la mise en œuvre d'une itération il faut $7M$ opérations, où $M = (N_1 - 1)(N_2 - 1)$ est le nombre d'inconnues dans le problème.

Pour l'exemple traité, l'opérateur B dans l'espace de dimension finie H des fonctions de mailles données sur ω à produit scalaire $(u, v) = \sum_{x\in\omega} u(x)\, v(x)\, h_1h_2$, $u, v \in H$ se définit de la façon suivante:

$$By = (\mathscr{D} + L)\, y = \left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right) \overset{\circ}{y}(i, j) - \frac{1}{h_1^2}\, \overset{\circ}{y}(i-1, j) -$$

$$- \frac{1}{h_2^2}\, \overset{\circ}{y}(i, j-1) = \left(\frac{1}{h_1^2} + \frac{1}{h_2^2}\right) \overset{\circ}{y} + \sum_{\alpha=1}^{2} \frac{1}{h_\alpha}\, \overset{\circ}{y}_{\bar{x}_\alpha},$$

où $y \in H$, $\overset{\circ}{y} \in \overset{\circ}{H}$ et $\overset{\circ}{y}(x) = y(x)$ pour $x \in \omega$. $\overset{\circ}{H}$ est ici un ensemble des fonctions de mailles données sur $\bar{\omega}$ et s'annulant sur γ.

Passons maintenant à la *méthode de Seidel par blocs.* Si l'on désigne par $Y_j = (y(1, j), y(2, j), \ldots, y(N_1 - 1, j))$ le vecteur composé d'inconnues sur la ligne j du maillage, alors, comme il a été montré au § 1, ch. I, le problème de différences (6) peut être écrit sous forme d'un système triponctuel d'équations vectorielles:

$$\begin{aligned} &-Y_{j-1} + CY_j - Y_{j+1} = F_j, \quad j = 1, 2, \ldots, N_2 - 1, \\ &Y_0 = F_0, \quad Y_{N_2} = F_{N_2}, \end{aligned} \tag{7}$$

où C est la matrice carrée tridiagonale de dimension $(N_1 - 1) \times \times (N_1 - 1)$, définie de la façon suivante:

$$(CY_j)_i = (2y - h_2^2 y_{\bar{x}_1x_1})_{ij}, \quad y_{0j} = y_{N_1j} = 0.$$

Les seconds membres F_j se définissent par les formules

$$F_j = \Big(h_2^2\varphi(1, j) + \frac{h_2^2}{h_1^2}\, g(0, j), h_2^2\varphi(2, j), \ldots, h_2^2\varphi(N_1 - 2, j),$$

$$h_2^2\varphi(N_1 - 1, j) + \frac{h_2^2}{h_1^2}\, g(N_1, j)\Big) \quad \text{pour } j = 1, 2, \ldots, N_2 - 1,$$

$$F_j = (g(1, j), g(2, j), \ldots, g(N_1 - 1, j)) \quad \text{pour } j = 0, N_2.$$

La méthode de Seidel par blocs pour le système (7) prend la forme

$$\begin{aligned} &CY_j^{(k+1)} = Y_{j-1}^{(k+1)} + Y_{j+1}^{(k)} + F_j, \quad j = 1, 2, \ldots, N_2 - 1, \\ &Y_0^{(k)} = F_0, \quad Y_{N_2}^{(k)} = F_{N_2}, \quad k = 0, 1, \ldots, \end{aligned} \tag{8}$$

et pour trouver $Y_j^{(k+1)}$ il faut inverser la matrice tridiagonale C.

Si l'on distribue le schéma (8) entre les points du maillage, on obtient les formules suivantes:

$$- \frac{1}{h_1^2}\, y_{k+1}(i-1, j) + \left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right) y_{k+1}(i, j) - \frac{1}{h_1^2}\, y_{k+1}(i+1, j) =$$

$$= \frac{1}{h_2^2}\, y_{k+1}(i, j-1) + \frac{1}{h_2^2}\, y_k(i, j+1) + \varphi(i, j), \tag{9}$$

$$1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant j \leqslant N_2 - 1,$$

avec $y_k(x) = g(x)$, $x \in \gamma$ pour tout $k \geqslant 0$. Pour trouver y_{k+1} sur la j-ième ligne, il faut résoudre le problème aux limites triponctuel (9), dont le second membre est connu, par la méthode du balayage, par exemple, et placer la solution obtenue à l'endroit de y_k sur la j-ième ligne.

A la méthode de Seidel par blocs correspond l'opérateur B suivant:

$$By = \frac{1}{h_2^2} \overset{\circ}{y} + \frac{1}{h_2} \overset{\circ}{y}_{\bar{x}_2} + \overset{\circ}{y}_{\bar{x}_1 x_1}, \quad y \in H, \quad \overset{\circ}{y} \in \overset{\circ}{H}.$$

Supposons maintenant que sur le maillage $\bar{\omega}$ il s'agisse de rechercher la solution du problème discret de Dirichlet pour l'équation elliptique à coefficients variables

$$\begin{gathered} \Lambda y = \sum_{\alpha=1}^{2} (a_\alpha(x) y_{\bar{x}_\alpha})_{x_\alpha} - d(x) y = -\varphi(x), \quad x \in \omega, \\ y(x) = g(x), \quad x \in \gamma, \end{gathered} \tag{10}$$

$$0 < c_1 \leqslant a_\alpha(x) \leqslant c_2, \quad x \in \bar{\omega}, \quad \alpha = 1, 2, \quad 0 \leqslant d_1 \leqslant d(x) \leqslant d_2, \quad x \in \omega.$$

Pour le problème posé, la méthode ponctuelle de Seidel, une fois les inconnues ordonnées suivant les lignes du maillage, prend la forme suivante:

$$\begin{aligned} \left(\frac{a_1(i+1, j) + a_1(i, j)}{h_1^2} + \frac{a_2(i, j+1) + a_2(i, j)}{h_2^2} + d(i, j) \right) y_{k+1}(i, j) = \\ = \frac{a_1(i, j)}{h_1^2} y_{k+1}(i-1, j) + \frac{a_2(i, j)}{h_2^2} y_{k+1}(i, j-1) + \\ + \frac{a_1(i+1, j)}{h_1^2} y_k(i+1, j) + \frac{a_2(i, j+1)}{h_2^2} y_k(i, j+1) + \varphi(i, j) \end{aligned}$$

pour $i = 1, 2, \ldots, N_1 - 1$ et $j = 1, 2, \ldots, N_2 - 1$, avec $y_k(x) = g(x)$ pour $x \in \gamma$ avec tout $k \geqslant 0$.

L'opérateur B sous forme canonique de schéma itératif se définit pour l'exemple donné de la façon suivante:

$$\begin{aligned} By(x_{ij}) = \left(\frac{a_1(i+1, j)}{h_1^2} + \frac{a_2(i, j+1)}{h_2^2} + d(i, j) \right) \overset{\circ}{y}(i, j) + \\ + \frac{a_1(i, j)}{h_1} \overset{\circ}{y}_{\bar{x}_1} + \frac{a_2(i, j)}{h_2} \overset{\circ}{y}_{\bar{x}_2}, \quad y \in H, \quad \overset{\circ}{y} \in \overset{\circ}{H}, \end{aligned}$$

où l'espace H et l'ensemble $\overset{\circ}{H}$ ont été définis plus haut.

3. Conditions suffisantes de convergence. Formulons maintenant quelques conditions suffisantes de convergence de la méthode de Seidel. Il nous faut pour cela le théorème suivant.

Théorème 1. *Soit dans l'équation* (1) *l'opérateur A autoadjoint et défini positif dans H. Dans ce cas le processus itératif à deux couches*

$$B \frac{y_{k+1} - y_k}{\tau} + A y_k = f, \quad k = 0, 1, \ldots, \quad y_0 \in H, \quad \tau > 0, \tag{11}$$

converge dans H_A, *si l'opérateur* $B - 0.5\tau A$ *est défini positif dans* H, *autrement dit si la condition*

$$B > \frac{\tau}{2} A \tag{12}$$

est remplie.

En effet, de (11) on obtient pour l'erreur $z_k = y_k - u$ le problème suivant:

$$B \frac{z_{k+1} - z_k}{\tau} + Az_k = 0, \quad k = 0, 1, \ldots, \quad z_0 = y_0 - u. \tag{13}$$

Etablissons pour z_k l'identité énergétique fondamentale. Portons dans (13) z_k sous la forme $z_k = \frac{1}{2}(z_{k+1} + z_k) - \frac{\tau}{2}\left(\frac{z_{k+1} - z_k}{\tau}\right)$ et il vient

$$\left(B - \frac{\tau}{2} A\right) \frac{z_{k+1} - z_k}{\tau} + \frac{1}{2} A (z_{k+1} + z_k) = 0.$$

Multiplions scalairement le premier et le second membre de cette égalité par $2(z_{k+1} - z_k)$ et tenons compte de ce que si l'opérateur A est autoadjoint on a l'égalité $(A(z_{k+1} + z_k), z_{k+1} - z_k) = (Az_{k+1}, z_{k+1}) - (Az_k, z_k)$. Finalement, on obtient l'identité énergétique fondamentale

$$2\tau \left(\left(B - \frac{\tau}{2} A\right) \frac{z_{k+1} - z_k}{\tau}, \frac{z_{k+1} - z_k}{\tau}\right) + \| z_{k+1} \|_A^2 - \| z_k \|_A^2 = 0.$$

De là et des inégalités $B - 0{,}5\tau A > 0$, $\tau > 0$ il s'ensuit que $\| z_{k+1} \|_A^2 \leqslant \| z_k \|_A^2$, c'est-à-dire que la suite $\{\| z_k \|_A^2\}$ ne s'accroît pas, est bornée inférieurement par zéro et est convergente. Il s'ensuit alors de l'identité énergétique que

$$\lim_{k \to \infty} \left(\left(B - \frac{\tau}{2} A\right) \frac{z_{k+1} - z_k}{\tau}, \frac{z_{k+1} - z_k}{\tau}\right) = 0. \tag{14}$$

Ensuite, de l'inégalité $B - 0{,}5\,\tau A > 0$ il découle que $\| z_{k+1} - z_k \| \to 0$ pour $k \to \infty$. En notant que dans (13) $A^{1/2} z_k = -A^{-1/2} B (z_{k+1} - z_k)/\tau$, il vient $\| z_k \|_A^2 \leqslant \| A^{-1} \| \| B \|^2 \times \| z_{k+1} - z_k \|^2/\tau^2 \to 0$ pour $k \to \infty$.

Formulons la condition suffisante de la convergence de la méthode de Seidel.

Théorème 2. *Si l'opérateur* A *est autoadjoint et défini positif dans* H, *alors la méthode de Seidel* (4), (5) *converge dans* H_A.

En effet, il s'ensuit de (5) et du théorème 1 qu'il suffit de vérifier l'inégalité $\mathscr{D} + L - 0{,}5\, A > 0$. Or, comme $A = A^*$, on a dans (4) $U = L^*$ et

$$((\mathscr{D} + L - 0{,}5\, A)\, x, x) = 0{,}5\, ((\mathscr{D} + L - U)\, x, x) = 0{,}5\, (\mathscr{D} x, x).$$

Vu que A est un opérateur défini positif, on a pour la méthode ponc-

tuelle de Seidel $a_{ii} > 0$, $1 \leqslant i \leqslant M$, et pour la méthode de Seidel par blocs les matrices $a_{ii} = a^*_{ii} > 0$. Par conséquent, $\mathscr{D} = \mathscr{D}^* > 0$. Donc $\mathscr{D} + L - 0{,}5\ A > 0$.

Fournissons, sans démonstration, encore une condition de la convergence de la méthode de Seidel.

Théorème 3. *Si l'opérateur A est autoadjoint et non dégénéré, tandis que tous les $a_{ii} > 0$, alors, pour toute approximation initiale, la méthode de Seidel converge seulement et rien que seulement si A est un opérateur défini positif.*

Pour apprécier la vitesse de convergence de la méthode de Seidel, on se sert de différentes espèces d'hypothèses.

Par exemple, si est satisfaite la condition

$$\sum_{j \neq i} |a_{ij}| \leqslant q\, |a_{ii}|, \quad i = 1, 2, \ldots, M, \quad q < 1, \tag{15}$$

la méthode de Seidel converge à la vitesse de la progression géométrique avec dénominateur q et pour l'erreur z_n on a l'estimation $\| z_n \| \leqslant q^n \| z_0 \|$, où $\| z_n \| = \max\limits_{1 \leqslant i \leqslant M} |y_i^{(n)} - u_i|$.

En effet, de (3) on obtient pour l'erreur $z_i^{(k)} = y_i^{(k)} - u_i$ l'équation homogène

$$a_{ii} z_i^{(k+1)} = -\sum_{j=1}^{i-1} a_{ij} z_j^{(k+1)} - \sum_{j=i+1}^{M} a_{ij} z_j^{(k)}.$$

De là on obtient

$$| z_i^{(k+1)} | \leqslant \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| | z_j^{(k+1)} | + \sum_{j=i+1}^{M} \left| \frac{a_{ij}}{a_{ii}} \right| | z_j^{(k)} | \leqslant$$
$$\leqslant \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \| z_{k+1} \| + \sum_{j=i+1}^{M} \left| \frac{a_{ij}}{a_{ii}} \right| \| z_k \|. \tag{16}$$

De (15), il vient

$$\sum_{j=i+1}^{M} \left| \frac{a_{ij}}{a_{ii}} \right| \leqslant q - \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \leqslant q \left(1 - \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \right).$$

En portant cette estimation dans (16), on obtient l'inégalité suivante:

$$| z_i^{(k+1)} | \leqslant \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \| z_{k+1} \| + q \left(1 - \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \right) \| z_k \|. \tag{17}$$

Supposons que $\max |z_i^{(k+1)}|$ est atteint pour un certain $i = i_0$, de sorte que $\|z_{k+1}\| = |z_{i_0}^{(k+1)}|$. De (17), pour $i = i_0$, on obtient

$$\left(1 - \sum_{j=1}^{i_0-1} \left|\frac{a_{i_0 j}}{a_{i_0 i_0}}\right|\right) \|z_{k+1}\| \leqslant q \left(1 - \sum_{j=1}^{i_0-1} \left|\frac{a_{i_0 j}}{a_{i_0 i_0}}\right|\right) \|z_k\|;$$

de là s'ensuit l'estimation $\|z_{k+1}\| \leqslant q \|z_k\| \leqslant \ldots \leqslant q^{k+1} \|z_0\|$. La proposition est démontrée.

La condition (15) signifie que A est une matrice à dominance diagonale. Pour les exemples d'application de la méthode de Seidel fournis au point 2 la condition (15) n'est pas remplie ($q = 1$). Dans ces exemples l'opérateur A est autoadjoint et défini positif dans H. Aussi, en vertu du théorème 2, ne peut-on qu'affirmer que la méthode converge dans H_A. L'estimation de la vitesse de convergence dans H_A sera fournie plus loin, après l'examen du schéma général des méthodes itératives triangulaires.

§ 2. Méthode de surrelaxation

1. Schéma itératif. Conditions suffisantes de convergence. Pour accélérer la convergence de la méthode de Seidel, on la modifie en introduisant dans le schéma itératif le paramètre d'itération ω, de sorte que

$$(\mathscr{D} + \omega L) \frac{y_{k+1} - y_k}{\omega} + A y_k = f, \quad k = 0, 1, \ldots, \quad y_0 \in H, \qquad (1)$$

où, comme auparavant, la matrice A se présente sous forme de somme

$$A = \mathscr{D} + L + U. \qquad (2)$$

La méthode de Seidel correspond à la valeur de $\omega = 1$.

En confrontant (1) avec la forme canonique des schémas itératifs à deux couches, on trouve que

$$B = \mathscr{D} + \omega L, \quad \tau_k \equiv \omega.$$

Comme pour la méthode de Seidel, dans la méthode étudiée à l'opérateur B correspond la matrice triangulaire inférieure, de sorte que l'introduction du paramètre ω ne nous sort pas de la classe des méthodes itératives triangulaires. Ce qui s'ajoute c'est le problème du choix du paramètre ω.

Si l'on répartit le schéma itératif (1) suivant les composantes du vecteur y_{k+1}, on obtient alors la formule suivante:

$$a_{ii} y_i^{(k+1)} = (1 - \omega) a_{ii} y_i^{(k)} - \omega \sum_{j=1}^{i-1} a_{ij} y_j^{(k+1)} - \omega \sum_{j=i+1}^{M} a_{ij} y_j^{(k)} + \omega f_i \qquad (3)$$

pour $i = 1, 2, \ldots, M$ ($y_i^{(k+1)}$ trouvé se place à l'endroit de $y_i^{(k)}$). La réalisation d'une itération s'effectue à peu près avec la même dépense en opérations arithmétiques que pour la méthode de Seidel.

La méthode itérative (1) pour $\omega > 1$ est appelée *méthode de surrelaxation*, pour $\omega = 1$ *de pleine relaxation* et pour $\omega < 1$ *de sousrelaxation*.

On a démontré au § 1 que la méthode de Seidel converge dans H_A au cas d'un opérateur A autoadjoint et défini positif. Pour la convergence de la méthode de relaxation, outre ces conditions, on oblige le paramètre d'itération ω de satisfaire à une condition supplémentaire. Formulons les conditions suffisantes pour la convergence de la méthode de relaxation.

Théorème 4. *Si l'opérateur A est autoadjoint et défini positif dans H, tandis que le paramètre ω satisfait à la condition $0 < \omega < 2$, la méthode de relaxation* (1) *converge alors dans H_A.*

En effet, du théorème 1 il s'ensuit qu'il suffit de s'assurer de la satisfaction de l'inégalité $\mathscr{D} + \omega L > 0{,}5\,\omega A$ pour $\omega > 0$. Comme $A = A^* > 0$, l'opérateur $\mathscr{D}$ est autoadjoint et défini positif dans H et $U = L^*$. Donc, en utilisant l'égalité (14), § 1, il vient

$$((\mathscr{D} + \omega L)\,x,\ x) = (1 - 0{,}5\,\omega)\ (\mathscr{D}\,x,\ x) + 0{,}5\omega\ ((\mathscr{D} + 2L)x,\ x) = \\ = (1 - 0{,}5\,\omega)\ (\mathscr{D}x,\ x) + 0{,}5\omega\ (Ax,\ x).$$

Avec $\omega < 2$ on en déduit la proposition du théorème.

Remarque. Le théorème 4 se vérifie aussi bien pour la méthode de relaxation ponctuelle, quand dans (3) a_{ij} sont des nombres, que pour la méthode de relaxation par blocs ou vectorielle, quand dans (3) a_{ij} représentent des matrices de dimensions correspondantes.

2. Position du problème de choix du paramètre d'itération. Le théorème 4 fournit des conditions suffisantes de convergence pour la méthode de relaxation en laissant irrésolu le problème du choix optimal du paramètre ω. La singularité du processus itératif étudié (1) consiste dans le fait que le paramètre d'itération ω entre dans l'opérateur $B = \mathscr{D} + \omega L$, qui est un opérateur non autoadjoint dans H. On a déjà eu affaire au cas d'un opérateur non autoadjoint dans le § 4, ch. VI, où on a étudié la méthode itérative simple dont le paramètre d'itération a été choisi sur la base des conditions variées, par exemple, sur la base de la condition du minimum de la norme de l'opérateur de passage d'une itération à l'autre. Dans le cas concerné il s'agit de tenir compte de la singularité du schéma itératif mentionnée plus haut. On procédera au choix du paramètre ω sur la base de la condition du minimum de la norme dans H_A de l'opérateur de passage d'une itération à l'autre au § 3 de ce chapitre, où on examinera le schéma général des méthodes itératives triangulaires. Au point présent le paramètre ω pour la méthode de relaxation sera choisi sur

la base de la condition du minimum du rayon spectral de l'opérateur de passage d'une itération à l'autre.

Rappelons la définition du rayon spectral de l'opérateur

$$\rho(S) = \lim_{n\to\infty} \sqrt[n]{\| S^n \|} = \max_k |\lambda_k|, \qquad (4)$$

où λ_k est la valeur propre de l'opérateur S. Le rayon spectral possède les propriétés suivantes:

$$\rho(S^n) = \rho^n(S), \quad \rho(S) \leqslant \| S \| \qquad (5)$$

et $\rho(S) = \| S \|$, si S est un opérateur autoadjoint dans H. A partir de (5), pour un opérateur S quelconque, on obtient $\rho^n(S) = \rho(S^n) \leqslant \leqslant \| S^n \|$. D'autre part, de (4), pour un n suffisamment grand, on aura $\rho^n(S) \approx \| S^n \|$.

Passons maintenant au problème du choix optimal du paramètre ω pour le schéma itératif (1). Posons d'abord le problème de l'erreur $z_k = y_k - u$. De (1), il vient

$$(\mathscr{D} + \omega L)\frac{z_{k+1} - z_k}{\omega} + Az_k = 0, \quad k = 0, 1, \ldots, \quad z_0 = y_0 - u$$

ou

$$z_{k+1} = Sz_k, \quad k = 0, 1, \ldots, \quad S = E - \omega(\mathscr{D} + \omega L)^{-1}A. \qquad (6)$$

En utilisant (6), exprimons z_n au moyen de z_0:

$$z_n = S^n z_0, \quad \| z_n \| \leqslant \| S^n \| \| z_0 \|. \qquad (7)$$

L'opérateur S est un opérateur non autoadjoint dans H, qui dépend du paramètre ω. Le problème du choix optimal du paramètre ω sera formulé de la façon suivante: chercher ω à partir de la condition du minimum du rayon spectral de l'opérateur S.

Il faut noter que l'on n'a pas minimisé la norme de l'opérateur résolvant S^n, comme il aurait fallu le faire en vertu de l'estimation (7), par contre, on minimise le rayon spectral $\rho(S)$ de l'opérateur de passage S pour lequel on a l'estimation $\rho^n(S) \leqslant \| S^n \|$. Toutefois, en vertu de l'égalité approchée $\rho^n(S) \approx \| S^n \|$, on peut s'attendre pour un n suffisamment grand à ce que le choix indiqué de ω s'avère heureux.

La résolution du problème formulé plus haut est une entreprise compliquée, mais en formulant quelques hypothèses complémentaires sur l'opérateur A ce problème peut être résolu avec succès.

H y p o t h è s e 1. L'opérateur A est autoadjoint et défini positif dans H $(U = L^*,\ \mathscr{D} = \mathscr{D}^* > 0)$.

H y p o t h è s e 2. L'opérateur A est tel que pour tout z complexe différent de zéro les valeurs propres μ du problème généralisé sur les valeurs propres $\left(zL + \frac{1}{z}U\right)x = \mu\mathscr{D}x = 0$ sont indépendantes de z.

En utilisant ces hypothèses, démontrons la proposition suivante qui nous sera utile dans la suite.

L e m m e 1. *Si l'opérateur A vérifie les hypothèses* 1 *et* 2, *alors toutes les valeurs propres du problème*

$$Ax - \lambda \mathscr{D} x = 0 \tag{8}$$

sont réelles, positives et, si λ est une valeur propre, $2 - \lambda$ *est aussi une valeur propre.*

En effet, la positivité et la nature réelle des valeurs propres λ se déduisent du fait que l'opérateur A est autoadjoint et défini positif. Ensuite, supposons que λ est la valeur propre du problème (8), c'est-à-dire que

$$Ax - \lambda \mathscr{D} x = (L + U)\, x - (\lambda - 1)\, \mathscr{D} x = 0, \quad x \neq 0.$$

En vertu de l'hypothèse 2 on aura l'égalité

$$(-L - U)\, y - (\lambda - 1)\, \mathscr{D} y = 0 \quad \text{ou} \quad Ay - (2 - \lambda)\, \mathscr{D} y = 0.$$

De là s'ensuit l'assertion du lemme.

Passons maintenant à la résolution du problème sur le choix optimal du paramètre ω. Pour cela il faut apprécier le rayon spectral de l'opérateur de passage $S = E - \omega\, (\mathscr{D} + \omega L)^{-1} A$, c'est-à-dire apprécier les valeurs propres μ de l'opérateur S:

$$Sx - \mu x = 0. \tag{9}$$

Posons que les hypothèses 1 et 2 sont vérifiées. Le lemme suivant établit le rapport entre les valeurs propres μ du problème (9) et les valeurs propres λ du problème (8).

L e m m e 2. *Pour $\omega \neq 1$ les valeurs propres du problème* (8) *et* (9) *sont liées par la relation*

$$(\mu + \omega - 1)^2 = \omega^2 \mu\, (1 - \lambda)^2. \tag{10}$$

En effet, soient μ et λ les valeurs propres du problème (9) et (8). De la définition de l'opérateur S et du développement $A = \mathscr{D} + L + U$ il s'ensuit que (9) peut être écrit sous la forme

$$\frac{1-\mu-\omega}{\omega}\, \mathscr{D} x - (\mu L + U)\, x = 0, \quad x \neq 0. \tag{11}$$

Montrons d'abord que pour $\omega \neq 1$ tous les μ sont différents de zéro. En effet, posons que $\mu = 0$. Alors (11) prend la forme

$$\frac{1-\omega}{\omega}\, \mathscr{D} x - Ux = 0.$$

Vu que U est une matrice triangulaire supérieure, tandis que $\mathscr{D}$ est une matrice diagonale (diagonale par blocs) qui est définie positive en vertu de l'hypothèse 1, cette dernière égalité peut se vérifier pour $x \neq 0$ seulement et rien que seulement si $\omega = 1$. On a donc abouti à une contradiction en supposant qu'on a $\mu = 0$ pour $\omega \neq 1$.

En divisant le premier et le second membre de (11) par $\sqrt{\mu}$, il vient

$$\frac{1-\mu-\omega}{\omega \sqrt{\mu}}\, \mathscr{D} x - \left(\sqrt{\mu}\, L + \frac{1}{\sqrt{\mu}}\, U \right) x = 0.$$

De là, en raison de l'hypothèse 2, on obtient

$$\frac{1-\mu-\omega}{\omega \sqrt{\mu}}\, \mathscr{D} y - (L + U)\, y = 0$$

ou

$$Ay - \left(1 + \frac{1-\mu-\omega}{\omega\sqrt{\mu}}\right)\mathscr{D}y = 0.$$

En comparant cette égalité à (8), on obtient la relation

$$\frac{\mu+\omega-1}{\omega\sqrt{\mu}} = 1-\lambda.$$

Ainsi s'achève la démonstration du lemme 2.

Remarque. Lors de la démonstration du lemme 2 on n'a pas utilisé le fait que l'opérateur A était autoadjoint. La relation (10) se vérifie également pour le cas de tout opérateur A non autoadjoint si l'opérateur $\mathscr{D}$ est non dégénéré.

Il s'ensuit du lemme 1 que les valeurs propres λ se placent sur l'axe réel de symétrie par rapport au point $\lambda = 1$, de plus, $\lambda \in [\lambda_{\min}, 2 - \lambda_{\min}]$, $\lambda_{\min} > 0$. Aussi obtient-on du lemme 2 que pour $\omega \neq 1$ à chaque $\lambda_i = 1$ correspond $\mu_i = 1 - \omega$, à chaque couple λ_i et $2 - \lambda_i$ correspond un couple de valeurs non nulles de μ_i obtenues en résolvant l'équation (10) avec $\lambda = \lambda_i$. Par conséquent, il est possible de trouver tous les μ_i, ces derniers étant les racines de l'équation quadratique (10), dans laquelle en guise de λ on prend tous les λ_i se disposant sur le tronçon $[\lambda_{\min}, 1]$.

3. Appréciation du rayon spectral. Cherchons maintenant la valeur optimale du paramètre ω et apprécions le rayon spectral de l'opérateur S. Pour cela étudions l'équation (10):

$$\mu^2 + [2(\omega-1) - \omega^2(1-\lambda)^2]\,\mu + (\omega-1)^2 = 0, \qquad (12)$$

où $\lambda_{\min} \leqslant \lambda \leqslant 1$ et $0 < \omega < 2$.

En résolvant l'équation (12), on obtient deux racines

$$\mu_1(\lambda, \omega) = \left(\frac{\omega(1-\lambda) + \sqrt{\omega^2(1-\lambda)^2 - 4(\omega-1)}}{2}\right)^2,$$
$$\mu_2(\lambda, \omega) = \left(\frac{\omega(1-\lambda) - \sqrt{\omega^2(1-\lambda)^2 - 4(\omega-1)}}{2}\right)^2. \qquad (13)$$

L'examen du discriminant de l'équation (12) permet de constater que pour $\omega > \omega_0 > 1$, où

$$\omega_0 = \frac{2}{1+\sqrt{\lambda_{\min}(2-\lambda_{\min})}} \in (1, 2), \qquad (14)$$

les racines μ_1 et μ_2 pour tout $\lambda \in [\lambda_{\min}, 1]$ sont complexes, avec $|\mu_1| = |\mu_2| = \omega - 1$. Aussi le rayon spectral de l'opérateur S pour $\omega > \omega_0$ est-il égal à $\rho(S) = \omega - 1$ et croît en ω. Si $\omega = \omega_0$, on a

$$\mu_1(\lambda_{\min}, \omega_0) = \mu_2(\lambda_{\min}, \omega_0) = \omega_0 - 1,$$

et pour $\lambda_{\min} < \lambda \leqslant 1$ les racines μ_1 et μ_2 redeviendront complexes avec $|\mu_1| = |\mu_2| = \omega_0 - 1$. Par conséquent, dans le domaine $\omega \geqslant \omega_0$ la valeur optimale de $\omega = \omega_0$ est celle, à laquelle correspond $\rho(S) = \omega_0 - 1$.

Supposons maintenant que $1 < \omega < \omega_0$. Etudions comment se comportent les racines μ_1 et μ_2 définies par la formule (13) comme étant des fonctions de la variable λ pour un ω fixé.

Si λ appartient au segment $[\lambda_{\min}, \lambda_0]$,

$$\lambda_{\min} \leqslant \lambda \leqslant \lambda_0 = 1 - 2\,\frac{\sqrt{\omega-1}}{\omega} < 1,$$

le discriminant $\omega^2(1-\lambda)^2 - 4(\omega-1)^2$ est alors non négatif et, par suite, les racines μ_1 et μ_2 sont réelles, la racine maximale étant la racine μ_1.

Montrons que $\mu_1(\lambda, \omega)$ est une fonction décroissante de λ sur le tronçon $[\lambda_{\min}, \lambda_0]$. En effet, en dérivant (12) en λ et compte tenu de (13), il vient

$$\frac{\partial \mu_1}{\partial \lambda} = -\frac{2\omega\mu_1}{\sqrt{\omega^2(1-\lambda)^2 - 4(\omega-1)}} < 0.$$

Par conséquent, la racine $\mu_1(\lambda, \omega)$ pour $1 < \omega < \omega_0$ décroît avec la variation de λ de $\lambda_{\min}$ à λ_0 et prend les valeurs suivantes:

$$\mu_1(\lambda_{\min}, \omega) = \left(\frac{\omega(1-\lambda_{\min}) + \sqrt{\omega^2(1-\lambda_{\min})^2 - 4(\omega-1)}}{2}\right)^2,$$

$$\mu_1(\lambda_0, \omega) = \omega - 1.$$

Ensuite, si λ varie de λ_0 à 1 les racines μ_1 et μ_2 sont complexes et égales en module: $|\mu_1| = |\mu_2| = \omega - 1$. Donc si $1 < \omega < \omega_0$, alors

$$\rho(S) = \mu_1(\lambda_{\min}, \omega) = \left(\frac{\omega(1-\lambda_{\min}) + \sqrt{\omega^2(1-\lambda_{\min})^2 - 4(\omega-1)}}{2}\right)^2. \qquad (15)$$

Si $\omega < 1$, toutes les racines de l'équation (12) sont réelles, la racine maximale étant la racine μ_1 dont les valeurs décroissent avec la variation de λ de $\lambda_{\min}$ à 1. Donc pour $\omega < 1$ le rayon spectral de l'opérateur S se définit par la formule (15). Vu que pour $\omega = 1$ les μ_k non nuls vérifient l'équation (12), l'équation (15) est vraie également pour $\omega = 1$.

Bref, si $0 < \omega < \omega_0$, le rayon spectral de l'opérateur S se définit par la formule (15). Montrons que $\mu_1(\lambda_{\min}, \omega)$ décroît en ω dans l'intervalle $0 < \omega < \omega_0$.

En effet, puisque pour $\omega < \omega_0$ la racine μ_1 décroît en λ pour $\lambda \leqslant \lambda_0$, tandis que $\mu_1(0, \omega) = 1$, alors $\mu_1(\lambda_{\min}, \omega) < 1$.

Ensuite, de (15) on obtient

$$\frac{\partial \mu_1(\lambda_{\min}, \omega)}{\partial \omega} = \sqrt{\mu_1}\left(1 - \lambda_{\min} + \frac{\omega(1-\lambda_{\min})^2 - 2}{\sqrt{\omega^2(1-\lambda_{\min})^2 - 4(\omega-1)}}\right) = \frac{\sqrt{\mu_1}}{\omega} \times$$

$$\times \frac{[\omega^2(1-\lambda_{\min})^2 - 2(\omega-1) + (1-\lambda_{\min})\,\omega\sqrt{\omega^2(1-\lambda_{\min})^2 - 4(\omega-1)} - 2]}{\sqrt{\omega^2(1-\lambda_{\min})^2 - 4(\omega-1)}}.$$

En y portant (13), on obtient finalement

$$\frac{\partial \mu_1}{\partial \omega} = \frac{2\sqrt{\mu_1}(\mu_1 - 1)}{\omega\sqrt{\omega^2(1-\lambda_{\min})^2 - 4(\omega-1)}} < 0.$$

La proposition est démontrée. Par conséquent, dans le domaine $\omega \leqslant \omega_0$ la valeur optimale est la valeur $\omega = \omega_0$ à laquelle correspond

$$\rho(S) = \omega_0 - 1 = \frac{1 - \sqrt{\lambda_{\min}(2-\lambda_{\min})}}{1 + \sqrt{\lambda_{\min}(2-\lambda_{\min})}} = \left(\frac{1-\sqrt{\eta}}{1+\sqrt{\eta}}\right)^2, \qquad \eta = \frac{\lambda_{\min}}{2-\lambda_{\min}}.$$

Notons qu'il s'ensuit des études menées plus haut que si δ est l'estimation de $\lambda_{\min}$ par le bas, c'est-à-dire si $\delta \leqslant \lambda_{\min}$, tandis que ω est choisi suivant la formule (14) avec la substitution de δ à $\lambda_{\min}$, on a $\omega_0 \leqslant \omega$, $\rho(S) \leqslant \left(\frac{1-\sqrt{\eta}}{1+\sqrt{\eta}}\right)^2$, $\eta = \frac{\delta}{2-\delta}$.

On a ainsi démontré le théorème suivant.

Théorème 5. *Supposons vérifiées les hypothèses* 1 *et* 2, δ *étant la constante de l'inégalité*

$$\delta\mathcal{D} \leqslant A, \quad \delta > 0. \tag{16}$$

Alors pour le rayon spectral de l'opérateur de passage S du schéma itératif (1) *avec la valeur optimale du paramètre* ω,

$$\omega = \omega_0 = \frac{2}{1+\sqrt{\delta(2-\delta)}}, \tag{17}$$

se vérifie l'estimation

$$\rho(S) \leqslant \left(\frac{1-\sqrt{\eta}}{1+\sqrt{\eta}}\right)^2, \quad \eta = \frac{\delta}{2-\delta}, \tag{18}$$

de plus, si dans (16) *on obtient une égalité, cette égalité s'observe également dans la formule* (18).

La méthode itérative (1), (17) est la *méthode de surrelaxation*, car $\omega_0 > 1$.

4. Problème discret de Dirichlet pour l'équation de Poisson dans un rectangle. Examinons comment s'applique la méthode de surrelaxation à la recherche de la solution approchée du problème discret de Dirichlet pour l'équation de Poisson donné sur un maillage rectangle $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G},\ 0 \leqslant i \leqslant N_1,\ 0 \leqslant j \leqslant N_2,\ h_\alpha = l_\alpha/N_\alpha,\ \alpha = 1, 2\}$ dans le rectangle $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha,\ \alpha = 1, 2\}$:

$$\Lambda y = \sum_{\alpha=1}^{2} y_{\bar{x}_\alpha x_\alpha} = -\varphi(x), \quad x \in \omega, \quad y(x) = g(x), \quad x \in \gamma. \tag{19}$$

L'opérateur A dans l'espace H des fonctions de mailles associées à ω avec produit scalaire

$$(u, v) = \sum_{x \in \omega} u(x)\, v(x)\, h_1 h_2$$

se définit de la façon ordinaire: $Ay = -\Lambda \overset{\circ}{y}$, $y \in H$, $\overset{\circ}{y} \in \overset{\circ}{H}$. Comme on le sait déjà, l'opérateur A correspondant au problème (19) est autoadjoint et défini positif dans H. Par suite, l'hypothèse 1 est satisfaite.

Examinons d'abord la *méthode ponctuelle de surrelaxation*. Si les inconnues sont ordonnées suivant les lignes du maillage ω, le schéma

aux différences (19) peut alors être écrit sous forme d'un système d'équations algébriques suivant:

$$-\frac{1}{h_1^2}y(i-1,\ j)-\frac{1}{h_2^2}y(i,\ j-1)+\left(\frac{2}{h_1^2}+\frac{2}{h_2^2}\right)y(i,\ j)-$$
$$-\frac{1}{h_1^2}y(i+1,\ j)-\frac{1}{h_2^2}y(i,\ j+1)=\varphi(i,\ j)$$

pour $i=1,\ 2,\ \ldots,\ N_1-1,\ j=1,\ 2,\ \ldots,\ N_2-1$ et $y(x)=$ $=g(x),\ x\in\gamma$.

A cette écriture de l'opérateur A correspond la représentation de A sous forme de somme $A=\mathscr{D}+L+U$, où

$$\mathscr{D}y=\left(\frac{2}{h_1^2}+\frac{2}{h_2^2}\right)y,$$
$$Ly(i,\ j)=-\frac{1}{h_1^2}\mathring{y}(i-1,\ j)-\frac{1}{h_2^2}\mathring{y}(i,\ j-1),$$
$$Uy(i,\ j)=-\frac{1}{h_1^2}\mathring{y}(i+1,\ j)-\frac{1}{h_2^2}\mathring{y}(i,\ j+1).$$

Pour le système étudié, la méthode ponctuelle de surrelaxation, conformément à la formule (3), prendra la forme suivante:

$$\left(\frac{2}{h_1^2}+\frac{2}{h_2^2}\right)y_{k+1}(i,\ j)=(1-\omega)\left(\frac{2}{h_1^2}+\frac{2}{h_2^2}\right)y_k(i,\ j)+$$
$$+\omega\left[\frac{1}{h_1^2}y_{k+1}(i-1,\ j)+\frac{1}{h_2^2}y_{k+1}(i,\ j-1)+\right.$$
$$\left.+\frac{1}{h_1^2}y_k(i+1,\ j)+\frac{1}{h_2^2}y_k(i,\ j+1)+\varphi(i,\ j)\right]$$

pour $i=1,\ 2,\ \ldots,\ N_1-1,\ j=1,\ 2,\ \ldots,\ N_2-1$, avec $y_k(x)=$ $=g(x)$ pour $x\in\gamma$ avec tout $k\geqslant 0$.

Les calculs, comme dans la méthode de Seidel, débutent avec le point $i=1,\ j=1$ et se poursuivent suivant les lignes ou suivant les colonnes du maillage ω. $y_{k+1}(i,\ j)$ ainsi trouvé se place à l'endroit de $y_k(i,\ j)$.

Démontrons maintenant que pour le paramètre considéré l'hypothèse 2 est vérifiée. Pour cela il faut montrer que pour tout z complexe différant de zéro les valeurs propres μ du problème

$$z\left(\frac{1}{h_1^2}y(i-1,\ j)+\frac{1}{h_2^2}y(i,\ j-1)\right)+$$
$$+\frac{1}{z}\left(\frac{1}{h_1^2}y(i+1,\ j)+\frac{1}{h_2^2}y(i,\ j+1)\right)+$$
$$+\mu\left(\frac{2}{h_1^2}+\frac{2}{h_2^2}\right)y(i,\ j)=0,\quad 1\leqslant i\leqslant N_1-1,\quad 1\leqslant j\leqslant N_2-1,$$
$$y(x)=0,\quad x\in\gamma$$

ne dépendent pas de z.

En effet, en posant ici

$y(i, j) = z^{i+j} v(i, j), \quad 0 \leqslant i \leqslant N_1, \quad 0 \leqslant j \leqslant N_2,$

il vient

$$\frac{1}{h_1^2} v(i-1, j) + \frac{1}{h_2^2} v(i, j-1) + \frac{1}{h_1^2} v(i+1, j) + \frac{1}{h_2^2} v(i, j+1) + \\ + \mu \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) v(i, j) = 0,$$

$$1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant j \leqslant N_2 - 1, \quad v(x) = 0, \quad x \in \gamma.$$

Par conséquent, μ ne dépend pas de z.

Il reste à trouver la valeur optimale du paramètre ω. A cette fin il faut obtenir ou apprécier par le bas la valeur propre minimale du problème (8), qui, pour le cas considéré, s'écrit sous la forme

$$y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} + \lambda \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y = 0, \quad x \in \omega, \quad y(x) = 0, \quad x \in \gamma.$$

Comme les valeurs propres de l'opérateur de différences de Laplace $\mathring{\Lambda} y = y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2}$ sont connues

$$\mathring{\lambda}_k = \frac{4}{h_1^2} \sin^2 \frac{k_1 \pi h_1}{2l_1} + \frac{4}{h_2^2} \sin^2 \frac{k_2 \pi h_2}{2l_2}, \quad k_\alpha = 1, 1, \ldots, N_\alpha - 1,$$

il en résulte alors que

$$\lambda_k = \mathring{\lambda}_k \Big/ \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) = \frac{2h_2^2}{h_1^2 + h_2^2} \sin^2 \frac{k_1 \pi h_1}{2l_1} + \frac{2h_1^2}{h_1^2 + h_2^2} \sin^2 \frac{k_2 \pi h_2}{2l_2}.$$

Donc

$$\lambda_{\min} = \frac{2h_2^2}{h_1^2 + h_2^2} \sin^2 \frac{\pi h_1}{2l_1} + \frac{2h_1^2}{h_1^2 + h_2^2} \sin^2 \frac{\pi h_2}{2l_2},$$

et le paramètre ω_0 s'obtient suivant la formule (14). Dans le cas particulier, où $\bar{G}$ est un carré de côté l ($l_1 = l_2 = l$) et le maillage est carré ($N_1 = N_2 = N$), on a

$$\lambda_{\min} = 2 \sin^2 \frac{\pi}{2N}, \quad \omega_0 = \frac{2}{1 - \sin \frac{\pi}{N}}, \quad \eta = \operatorname{tg}^2 \frac{\pi}{2N},$$

$$\rho(S) = \frac{1 - \sin \frac{\pi}{N}}{1 + \sin \frac{\pi}{N}} \approx 1 - \frac{2\pi}{N}.$$

Notons que le rayon spectral de l'opérateur de passage correspondant à la méthode ponctuelle de Seidel est apprécié suivant la formule (15) dans laquelle il faut poser $\omega = 1$. On obtient alors $\rho(S) = (1 - \lambda_{\min})^2 = \cos^2 \frac{\pi}{N}$, résultat de beaucoup inférieur à celui de la méthode de surrelaxation.

Voyons à présent la méthode de surrelaxation par blocs. Si l'on réunit dans le bloc les inconnues $y(i, j)$ sur la j-ième ligne du maillage, il correspond alors à l'écriture par blocs de l'opérateur A la représentation suivante $A = \mathscr{D} + L + U$, où

$$\mathscr{D}y = -\frac{1}{h_1^2}\overset{\circ}{y}(i-1, j) + \left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right)\overset{\circ}{y}(i, j) - \frac{1}{h_1^2}\overset{\circ}{y}(i+1, j),$$

$$Ly(i, j) = -\frac{1}{h_2^2}\overset{\circ}{y}(i, j-1), \quad Uy(i, j) = -\frac{1}{h_2^2}\overset{\circ}{y}(i, j+1).$$

Les formules de calculs pour la méthode de surrelaxation par blocs ont pour expression

$$-\frac{1}{h_1^2} y_{k+1}(i-1, j) + \left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right) y_{k+1}(i, j) - \frac{1}{h_2^2} y_{k+1}(i+1, j) =$$
$$= (1-\omega)\left(-\frac{1}{h_1^2} y_k(i-1, j) + \left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right) y_k(i, j) - \frac{1}{h_2^2} y_k(i+1, j)\right) +$$
$$+ \omega\left(\frac{1}{h_2^2} y_{k+1}(i, j-1) + \frac{1}{h_2^2} y_k(i, j+1) + \varphi(i, j)\right),$$
$$1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant j \leqslant N_2 - 1,$$

avec $y_k(x) = g(x)$, $x \in \gamma$ pour tous les $k \geqslant 0$. Pour obtenir y_{k+1} sur la j-ième ligne il faut résoudre, par exemple par la méthode du balayage, le problème aux limites triponctuel.

Montrons que pour l'exemple considéré l'hypothèse 2 est vérifiée, c'est-à-dire que les valeurs propres μ du problème

$$z\frac{1}{h_2^2} y(i, j-1) + \frac{1}{z}\frac{1}{h_2^2} y(i, j+1) + \mu\left(-\frac{1}{h_1^2} y(i-1, j) + \right.$$
$$\left. + \left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right) y(i, j) - \frac{1}{h_2^2} y(i+1, j)\right) = 0,$$
$$1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant j \leqslant N_2 - 1, \quad y(x) = 0, \quad x \in \gamma$$

ne dépendent pas de z. Cela s'établit sans peine par substitution $y(i, j) = z^j v(i, j)$, $0 \leqslant i \leqslant N_1$, $0 \leqslant j \leqslant N_2$.

Cherchons à présent la valeur optimale du paramètre ω. Le problème correspondant (8) a la forme

$$y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} + \lambda\left(\frac{2}{h_2^2} y - x_{\bar{x}_1 x_1}\right) = 0, \quad x \in \omega, \quad y(x) = 0, \quad x \in \gamma. \quad (20)$$

On vérifie sans peine que les fonctions propres du problème (20) sont

$$y_k(x) = \sin\frac{k_1 \pi x_1}{l_1} \sin\frac{k_2 \pi x_2}{l_2}. \quad (21)$$

En portant (21) dans (20), il vient

$$\lambda_k = \frac{\overset{\circ}{\lambda}_{k_1} + \overset{\circ}{\lambda}_{k_2}}{\frac{2}{h} + \overset{\circ}{\lambda}_{k_1}}, \quad k_\alpha = 1, 2, \ldots, N_\alpha - 1, \quad k = (k_1, k_2),$$

où

$$\overset{\circ}{\lambda}_{k_\alpha}=\frac{4}{h_\alpha^2}\sin^2\frac{k_\alpha\pi h_\alpha}{2l_\alpha},\quad k_\alpha=1,\ 2,\ \ldots,\ N_\alpha-1,\quad \alpha=1,2.$$

De là on tire

$$\lambda_{\min}=\frac{2h_2^2\sin^2\dfrac{\pi h_1}{2l_1}+2h_1^2\sin^2\dfrac{\pi h_2}{2l_2}}{2h_2^2\sin^2\dfrac{\pi h_1}{2l_1}+h_1^2}.$$

Pour le cas particulier considéré plus haut on aura

$$\lambda_{\min}=\frac{4\sin^2\dfrac{\pi}{2N}}{1+2\sin^2\dfrac{\pi}{2N}},\quad \omega_0=\frac{2+4\sin^2\dfrac{\pi}{2N}}{\left(1+\sqrt{2}\sin\dfrac{\pi}{2N}\right)^2},$$

$$\eta=2\sin^2\frac{\pi}{2N},\quad \rho(S)=\left(\frac{1-\sqrt{2}\sin\dfrac{\pi}{2N}}{1+\sqrt{2}\sin\dfrac{\pi}{2N}}\right)^2\approx 1-2\sqrt{2}\,\frac{\pi}{N}.$$

En comparant les estimations du rayon spectral des méthodes de surrelaxation ponctuelle et par blocs, on constate que la méthode par blocs converge $\sqrt{2}$ fois plus vite que la méthode ponctuelle. Mais d'un autre côté, la méthode par blocs exige pour chaque itération une dépense supérieure en opérations arithmétiques que la méthode ponctuelle.

Pour conclure, donnons le nombre d'itérations exigées par la méthode ponctuelle de surrelaxation en fonction du nombre de nœuds N suivant une direction pour $\varepsilon = 10^{-4}$. En qualité de problème modèle prenons le schéma aux différences (19) associé au maillage carré avec $N_1 = N_2 = N$ et $\varphi(x) \equiv 0$, $g(x) \equiv 0$. L'approximation initiale $y_0(x)$ est choisie de la façon suivante: $y_0(x) = 1$, $x \in \omega$, $y_0(x) = 0$, $x \in \gamma$.

Le processus d'itérations s'achèvera au moment où la condition

$$\|z_n\|_A \leqslant \varepsilon \|z_0\|_A \qquad (22)$$

sera remplie.

Il s'ensuit de la théorie de la méthode que l'erreur z_n possède l'estimation $\|z_n\|_A \leqslant \|S^n\|_A \|z_0\|_A$ et comme le rayon spectral de l'opérateur est inférieur ou égal à la norme de l'opérateur, on a $\rho^n(S) \leqslant \|S^n\|_A$. La condition $\rho^n(S) \leqslant \varepsilon$ ne peut donc être utilisée pour l'estimation du nombre d'itérations exigé.

Donnons le nombre d'itérations n déduit de la condition (22), et, à titre de comparaison, cherchons le nombre d'itérations n^* qui

s'ensuit de l'inégalité $\rho^n(S) \leqslant \varepsilon$:

$$N = 32 \quad n = 65 \quad n^* = 47$$
$$N = 64 \quad n = 128 \quad n^* = 94$$
$$N = 128 \quad n = 257 \quad n^* = 187$$

La comparaison des nombres d'itérations de la méthode de surrelaxation et de la méthode explicite de Tchébychev, examinée pour le problème (19) au point 1, § 5, ch. VI, montre que la méthode de surrelaxation exige à peu près 1,6 fois moins d'itérations que la méthode explicite de Tchébychev. Le nombre d'opérations arithmétiques que coûte une itération est pratiquement le même pour ces deux méthodes.

5. Problème discret de Dirichlet pour l'équation elliptique à coefficients variables. Examinons maintenant l'application de la méthode de surrelaxation à la recherche de la solution approchée du problème discret de Dirichlet pour une équation à coefficients variables dans un rectangle

$$\begin{aligned} &\Lambda y = \sum_{\alpha=1}^{2} (a_\alpha(x) y_{\bar{x}_\alpha})_{x_\alpha} - d(x) y = -\varphi(x), \quad x \in \omega, \\ &y(x) = g(x), \quad x \in \gamma, \end{aligned} \tag{23}$$

en posant remplies les conditions suivantes:

$$\begin{aligned} &0 < c_1 \leqslant a_\alpha(x) \leqslant c_2, \quad x \in \bar{\omega}, \quad \alpha = 1, 2, \\ &0 \leqslant d_1 \leqslant d(x) \leqslant d_2, \quad x \in \omega. \end{aligned} \tag{24}$$

Pour le problème (23), la méthode ponctuelle de surrelaxation, les inconnues étant ordonnées suivant les lignes du maillage ω, est décrite par la formule

$$\begin{aligned} b(i, j)\, y_{k+1}(i, j) = (1-\omega)\, b(i, j)\, y_k(i, j) + \\ + \omega \Big[\frac{a_1(i, j)}{h_1^2} y_{k+1}(i-1, j) + \frac{a_2(i, j)}{h_2^2} y_{k+1}(i, j-1) + \\ + \frac{a_1(i+1, j)}{h_1^2} y_k(i+1, j) + \frac{a_2(i, j+1)}{h_2^2} y_k(i, j+1) + \varphi(i, j) \Big], \\ 1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant j \leqslant N_2 - 1, \end{aligned} \tag{25}$$

où

$$b(i, j) = \frac{a_1(i, j) + a_1(i+1, j)}{h_1^2} + \frac{a_2(i, j) + a_2(i, j+1)}{h_2^2} + d(i, j)$$

et $y_k(x) = g(x)$, $x \in \gamma$ pour tout $k \geqslant 0$.

Pour l'exemple considéré, les opérateurs $\mathscr{D}$, L et U se définissent de la façon suivante:

$$\mathscr{D}y = by,$$

$$Ly(i, j) = -\frac{a_1(i, j)}{h_1^2}\overset{\circ}{y}(i-1, j) - \frac{a_2(i, j)}{h_2^2}\overset{\bullet}{y}(i, j-1),$$

$$Uy(i, j) = -\frac{a_1(i+1, j)}{h_1^2}\overset{\circ}{y}(i+1, j) - \frac{a_2(i, j+1)}{h_2^2}\overset{\bullet}{y}(i, j+1).$$

Les hypothèses 1 et 2 se vérifient, la démonstration étant la même que pour l'exemple du point 4.

Pour obtenir le paramètre ω, il faut trouver l'estimation de la constante δ dans l'inégalité $A \geqslant \delta\mathscr{D}$. Ce problème a été résolu auparavant dans le point 3, § 5, ch. VI, où on a examiné la méthode implicite de Tchébychev du type le plus simple pour un problème de différences (23). Donnons l'estimation de δ:

$$\delta = \min_{0<x_2<l_2} \frac{1}{\varkappa_1(x_2)} + \min_{0<x_1<l_1} \frac{1}{\varkappa_2(x_1)},$$

où $\varkappa_\alpha(x_\beta) = \max\limits_{0<x_\alpha<l_\alpha} v^\alpha(x)$, $\beta = 3-\alpha$, $\alpha = 1, 2$, tandis que $v^\alpha(x)$ est la solution du problème aux limites triponctuel suivant:

$$\left(a_\alpha v^\alpha_{\bar{x}_\alpha}\right)_{x_\alpha} - \frac{1}{2}dv^\alpha = -b(x), \quad h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha,$$

$$v^\alpha(x) = 0, \quad x_\alpha = 0,\ l_\alpha,$$

$$h_\beta \leqslant x_\beta \leqslant l_\beta - h_\beta, \quad \beta = 3-\alpha, \quad \alpha = 1, 2.$$

Le paramètre d'itération ω s'obtient suivant la formule (17):

$$\omega = \omega_0 = \frac{2}{1+\sqrt{\delta(2-\delta)}}.$$

Pour comparer la méthode de surrelaxation décrite avec la méthode implicite de Tchébychev du type le plus simple étudiée au point 3, § 5, ch. VI, indiquons le nombre d'itérations exigées par la méthode de surrelaxation pour l'exemple modèle suivant. Supposons le schéma aux différences (23) donné sur un maillage carré avec $N_1 = N_2 = N$ et $\varphi(x) = 0$, $g(x) = 0$. Les coefficients $a_1(x)$, $a_2(x)$ et $d(x)$ seront choisis de la façon suivante:

$$a_1(x) = 1 + c[(x_1 - 0{,}5)^2 + (x_2 - 0{,}5)^2],$$

$$a_2(x) = 1 + c[0{,}5 - (x_1 - 0{,}5)^2 - (x_2 - 0{,}5)^2],$$

$$d(x) \equiv 0, \quad c > 0.$$

De plus, dans les inégalités (24) on a $c_1 = 1$, $c_2 = 1 + 0{,}5c$, $d_1 = d_2 = 0$. L'approximation initiale de la méthode itérative de surre-

laxation (25) sera choisie de la façon suivante: $y_0(x) = 1$, $x \in \omega$, $y_0(x) = 0$, $x \in \gamma$, le processus d'itérations s'achevant avec l'observation de la condition (22).

Donnons au tableau 9 le nombre d'itérations de la méthode de relaxation en fonction du rapport c_2/c_1 et du nombre de nœuds N suivant une direction pour $\varepsilon = 10^{-4}$. Pour le cas où $a_\alpha(x) \equiv 1$ et $d(x) \equiv 0$ le nombre d'itérations de la méthode de surrelaxation est donné au point 4 du présent paragraphe.

Tableau 9

c_1/c_2	2	8	32	128	512
$N=32$	65	81	95	96	98
$N=64$	129	164	192	193	195

Il s'ensuit du tableau 9 que le nombre d'itérations de la méthode de surrelaxation de l'exemple modèle est environ deux fois moindre que le nombre d'itérations de la méthode implicite de Tchébychev du type le plus simple. Vu que le nombre d'opérations arithmétiques dépensées pour chaque itération des méthodes mentionnées est le même, la méthode de surrelaxation s'avère deux fois plus efficiente par rapport à la méthode implicite de Tchébychev du type le plus simple.

§ 3. Méthodes triangulaires

1. Schéma itératif. Aux §§ 1, 2 on a étudié deux méthodes, la méthode de Seidel et la méthode de relaxation. Ces méthodes appartiennent à la classe des méthodes implicites à deux couches, à l'opérateur B desquelles correspond une matrice triangulaire ou triangulaire par blocs. Sous la forme canonique, le schéma itératif de ces méthodes a pour aspect:

$$(\mathscr{D} + \omega L)\frac{y_{k+1} - y_k}{\omega} + Ay_k = f, \quad k = 0, 1, \ldots, y_0 \in H, \qquad (1)$$

où $\mathscr{D}$ et L sont des opérateurs obtenus par développement de A en une somme de matrices diagonale, triangulaires inférieure et supérieure

$$A = \mathscr{D} + L + U. \qquad (2)$$

A la méthode de Seidel correspond la valeur du paramètre $\omega = 1$.

Au cas de l'opérateur A autoadjoint et défini positif dans H la condition suffisante de convergence dans H_A de la méthode itéra-

tive (1) prend la forme

$$0 < \omega < 2. \tag{3}$$

Au § 2 on a discuté la question du choix optimal du paramètre d'itération ω. En posant que les hypothèses 1 et 2 sont vérifiées et l'information à priori donnée sous forme de constante δ de l'inégalité

$$\delta \mathscr{D} \leqslant A, \quad \delta > 0, \tag{4}$$

on a démontré que la valeur optimale ω pour laquelle le rayon spectral de l'opérateur de passage S du schéma (1) est minimisé est définie par la formule

$$\omega = \omega_0 = \frac{2}{1+\sqrt{\delta(2-\delta)}}. \tag{5}$$

Aux points 4, 5 du § 2 on a considéré les exemples des problèmes pour lesquels les hypothèses 1 et 2 sont vérifiées. Ces hypothèses se vérifient également pour des problèmes plus compliqués, par exemple, pour le schéma aux différences pentaponctuel approximant sur un maillage irrégulier en un domaine arbitraire le problème de Dirichlet pour une équation elliptique à coefficients variables.

Il existe, toutefois, des exemples pour lesquels l'hypothèse 2 n'est pas vérifiée. S'y rapportent le problème discret de Dirichlet pour une équation elliptique aux dérivées mixtes, le problème discret de Dirichlet d'ordre de précision élevé, etc.

La non-universalité du procédé de choix du paramètre d'itération ω et l'absence d'estimations de la vitesse de convergence de la méthode en une norme quelconque constituent les principaux défauts de la théorie développée au § 2.

Dans le présent paragraphe on étudiera le schéma général des méthodes itératives triangulaires dont le paramètre d'itération ω est choisi sur la base de la condition de la minimisation dans H_A de la norme de l'opérateur de passage. On définira également l'estimation de la vitesse de convergence de la méthode dans H_A dans l'hypothèse de l'opérateur A autoadjoint et défini positif.

Commençons l'étude des méthodes triangulaires par la transformation du schéma itératif (1). Introduisons les opérateurs R_1 et R_2 de la façon suivante:

$$R_1 = \frac{1}{2}\mathscr{D} + L, \quad R_2 = \frac{1}{2}\mathscr{D} + U.$$

Le développement (2) prend alors la forme

$$A = R_1 + R_2, \tag{6}$$

et si A est autoadjoint dans H, les opérateurs R_1 et R_2 sont mutuellement autoadjoints

$$R_1 = R^*. \tag{7}$$

En portant $L = R_1 - \frac{1}{2}\mathscr{D}$ dans (1) et en posant

$$\tau = 2\omega/(2-\omega), \tag{8}$$

écrivons le schéma itératif (1) sous une forme équivalente

$$(\mathscr{D}+\tau R_1)\frac{y_{k+1}-y_k}{\tau}+Ay_k=f, \quad k=0,\ 1,\ \ldots,\ y_0\in H, \tag{9}$$

avec, en vertu de (3), (8), $\tau > 0$.

Le schéma (9) peut être considéré indépendamment du schéma (1). A savoir, admettons que l'opérateur A autoadjoint dans H est représenté suivant la formule (6) sous forme de somme d'opérateurs R_1 et R_2 mutuellement autoadjoints, tandis que $\mathscr{D}$ est un opérateur autoadjoint quelconque défini positif dans H. On appellera le schéma itératif (9) *forme canonique des méthodes itératives triangulaires.* On continuera d'appeler ces méthodes triangulaires même au cas où les matrices associées aux opérateurs R_1 et R_2 ne sont pas triangulaires et la matrice associée à l'opérateur $\mathscr{D}$ n'est pas une matrice diagonale.

Il s'ensuit du théorème 1 que pour un opérateur A défini positif la méthode itérative (9) pour $\tau > 0$ converge dans H_A. En effet, il suffit pour cela d'établir que l'inégalité $\mathscr{D} + \tau R_1 > 0{,}5\tau A$ est vraie. De (7) il vient

$$(Ax, x) = (R_1x, x) + (R_2x, x) = 2\,(R_1x, x) = 2\,(R_2x, x) \tag{10}$$

et, par conséquent,

$$((\mathscr{D} + \tau R_1)\,x,\ x) = (\mathscr{D}x,\ x) + 0{,}5\tau\,(Ax,\ x) > 0{,}5\tau\,(Ax,\ x),$$

ce qu'il fallait démontrer.

En conclusion, notons que dans le schéma (9) à la méthode de Seidel correspond la valeur $\tau = 2$, tandis qu'à la méthode de surrelaxation, la valeur $\tau = 2/\sqrt{\delta\,(2-\delta)}$.

2. Appréciation de la vitesse de convergence. Apprécions à présent la vitesse de convergence de la méthode itérative du schéma (9) dans H_A, en posant A autoadjoint et défini positif dans H.

Le passage dans (9) à l'erreur $z_k = y_k - u$ fournit un schéma homogène en z_k

$$B\frac{z_{k+1}-z_k}{\tau}+Az_k=0, \quad k=0,\ 1,\ \ldots,\ z_0=y_0-u, \quad B=\mathscr{D}+\tau R_1,$$

d'où on tire

$$z_{k+1} = Sz_k, \quad k = 0,\ 1,\ \ldots, \quad S = E - \tau B^{-1}A,$$
$$\| z_{k+1} \|_A \leqslant \| S \|_A \, \| z_k \|_A. \tag{11}$$

Apprécions la norme de l'opérateur de passage S dans H_A. De la définition de la norme de l'opérateur on obtient

$$\| S \|_A^2 = \sup_{x \neq 0} \frac{(ASx, Sx)}{(Ax, x)} =$$

$$= \sup_{x \neq 0} \left[1 - 2\tau \frac{(B^{-1}Ax, Ax)}{(Ax, x)} + \tau^2 \frac{(AB^{-1}Ax, B^{-1}Ax)}{(Ax, x)} \right]. \quad (12)$$

Transformons l'expression entre crochets. Utilisant (10) et la définition de l'opérateur B, il vient

$$(By, y) = (\mathscr{D}y, y) + \tau (R_1 y, y) = (\mathscr{D}y, y) + 0{,}5\tau (Ay, y).$$

De là on tire $\tau^2 (Ay, y) = 2\tau (By, y) - 2\tau (\mathscr{D}y, y)$ ou après substitution $y = B^{-1}Ax$

$$\tau^2 (AB^{-1}Ax, B^{-1}Ax) = 2\tau (B^{-1}Ax, Ax) - 2\tau (\mathscr{D}B^{-1}Ax, B^{-1}Ax).$$

En portant cette expression dans (12), on aura

$$\| S \|_A^2 = \sup_{x \neq 0} \left[1 - 2\tau \frac{(\mathscr{D}B^{-1}Ax, B^{-1}Ax)}{(Ax, x)} \right].$$

Poursuivons les transformations. En posant $x = (B^*)^{-1}\mathscr{D}^{1/2}y$, on obtient

$$\frac{(\mathscr{D}B^{-1}Ax, B^{-1}Ax)}{(Ax, x)} = \frac{(Cy, Cy)}{(Cy, y)}, \quad C = \mathscr{D}^{1/2}B^{-1}A (B^*)^{-1} \mathscr{D}^{1/2}.$$

Vu que l'opérateur C est autoadjoint et défini positif dans H, en posant $y = C^{-1/2}\mathscr{D}^{-1/2}B^*v$, on obtient

$$\frac{(\mathscr{D}B^{-1}Ax, B^{-1}Ax)}{(Ax, x)} = \frac{(Av, v)}{(B\mathscr{D}^{-1}B^*v, v)}.$$

Bref, on obtient finalement

$$\| S \|_A^2 = \sup_{v \neq 0} \left[1 - 2\tau \frac{(Av, v)}{(B\mathscr{D}^{-1}B^*v, v)} \right].$$

De là on tire, si γ_1 est une quantité de l'inégalité

$$\gamma_1 B\mathscr{D}^{-1}B^* \leqslant A, \quad (13)$$

que

$$\|S\|_A \leqslant (1 - 2\tau\gamma_1)^{1/2}. \quad (14)$$

Comme γ_1 dépend du paramètre τ, la valeur optimale de τ peut être obtenue en recherchant avec des hypothèses complémentaires sur les opérateurs $\mathscr{D}$, R_1 et R_2 l'expression de γ_1.

3. Choix du paramètre d'itération. Choisissons maintenant le paramètre τ. On aura besoin du

L e m m e 3. *Soient δ et Δ les constantes des inégalités*

$$\delta\mathscr{D} \leqslant A, \quad R_1\mathscr{D}^{-1}R_2 \leqslant \frac{\Delta}{4} A, \quad \delta > 0. \quad (15)$$

Alors dans l'inégalité (13)

$$\gamma_1 = \delta \Big/ \left(1 + \tau\delta + \tau^2 \frac{\delta\Delta}{4}\right). \tag{16}$$

En effet, comme $B^* = \mathscr{D} + \tau R_2$, on a

$$B\mathscr{D}^{-1}B^* = (\mathscr{D} + \tau R_1)\,\mathscr{D}^{-1}\,(\mathscr{D} + \tau R_2) = \mathscr{D} + \tau\,(R_1 + R_2) + \\ + \tau^2 R_1 \mathscr{D}^{-1} R_2 = \mathscr{D} + \tau A + \tau^2 R_1 \mathscr{D}^{-1} R_2.$$

En utilisant les hypothèses (15), on en tire

$$B\mathscr{D}^{-1}B^* \leqslant (1/\delta + \tau + \tau^2\Delta/4)\,A.$$

Le lemme est démontré.

Ainsi donc, si l'information à priori a la forme des constantes δ et Δ des inégalités (15), γ_1 est apprécié alors suivant la formule (16).

En portant (16) dans (14), il vient

$$\| S \|_A^2 \leqslant \varphi(\tau) = 1 - 2\tau\delta \Big/ \left(1 + \tau\delta + \tau^2 \frac{\delta\Delta}{4}\right).$$

Il ne reste qu'à minimiser la fonction $\varphi(\tau)$. En égalant la dérivée $\varphi'(\tau)$ à zéro, il vient

$$\varphi'(\tau) = \frac{2\delta\left(\tau^2 \frac{\delta\Delta}{4} - 1\right)}{\left(1 + \tau\delta + \tau^2 \frac{\delta\Delta}{4}\right)^2} = 0, \qquad \tau_0 = \frac{2}{\sqrt{\delta\Delta}}.$$

Étant donné que pour $\tau < \tau_0$ la dérivée $\varphi'(\tau) < 0$, tandis que pour $\tau > \tau_0$ la dérivée $\varphi'(\tau) > 0$, pour $\tau = \tau_0$ la fonction $\varphi(\tau)$ atteint un minimum égal à $\varphi(\tau_0) = (1 - \sqrt{\eta})/(1 + \sqrt{\eta})$, $\eta = \delta/\Delta$. On a ainsi démontré le théorème 6.

Théorème 6. *Soient A et $\mathscr{D}$ des opérateurs autoadjoints et définis positifs dans H, δ et Δ des constantes dans* (15). *La méthode itérative triangulaire* (9), (6) *pour $\tau = \tau_0 = 2/\sqrt{\delta\Delta}$ converge dans H_A, tandis que pour l'erreur z_n se vérifie l'estimation $\| z_n \|_A \leqslant \rho^n \| z_0 \|_A$. Pour le nombre d'itérations n se vérifie l'estimation $n \geqslant n_0(\varepsilon)$,*

$$n_0(\varepsilon) = \ln \varepsilon / \ln \rho,$$

où $\rho = \left(\frac{1 - \sqrt{\eta}}{1 + \sqrt{\eta}}\right)^{1/2}$, $\quad \eta = \frac{\delta}{\Delta}$.

4. Appréciation de la vitesse de convergence des méthodes de Seidel et de relaxation. Le théorème 6 démontré, il est possible d'obtenir l'estimation de la vitesse de convergence dans H_A des méthodes de Seidel et de surrelaxation étudiées plus haut. Au point 2 du § 1 et au point 4 du § 2 les méthodes mentionnées ont été appliquées à la recherche de la solution approchée du problème discret de Dirichlet

pour l'équation de Poisson sur un maillage rectangulaire

$$\bar{\omega} = \{x_{ij} = (ih_1, jh_2), \quad 0 \leqslant i \leqslant N_1, \quad 0 \leqslant j \leqslant N_2, \quad h_\alpha = l_\alpha / N_\alpha, \ \alpha = 1, 2\}$$

$$\Lambda y = y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x} = -\varphi(x), \ x \in \omega,$$

$$y(x) = g(x), \quad x \in \gamma.$$

Le schéma itératif de ces méthodes avait la forme (1), où

$$\mathscr{D} y = \left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right) \overset{\circ}{y},$$

$$Ly(i, j) = -\frac{1}{h_1^2} \overset{\bullet}{y}(i-1, j) - \frac{1}{h_2^2} \overset{\bullet}{y}(i, j-1),$$

$$Uy(i, j) = -\frac{1}{h_1^2} \overset{\bullet}{y}(i+1, j) - \frac{1}{h_2^2} \overset{\bullet}{y}(i, j+1).$$

Pour la méthode de Seidel $\omega = 1$, tandis que pour la méthode de surrelaxation ω se définissait suivant la formule (5), où δ de l'inégalité (4) était estimé ainsi:

$$\delta = \frac{2h_2}{h_1^2 + h^2} \sin^2 \frac{\pi h_1}{2l_1} + \frac{2h_1^2}{h_1^2 + h_2^2} \sin^2 \frac{\pi h_2}{2l_2}. \tag{17}$$

Réduisons le schéma (1) pour l'exemple considéré à la forme (9). Pour cela, définissons les opérateurs R_1 et R_2:

$$R_1 y = \left(\frac{1}{2} \mathscr{D} + L\right) y = \frac{1}{h_1} \overset{\bullet}{y}_{\bar{x}_1} + \frac{1}{h_2} \overset{\circ}{y}_{\bar{x}_2},$$

$$R_2 y = \left(\frac{1}{2} \mathscr{D} + U\right) y = -\frac{1}{h_1} \overset{\circ}{y}_{x_1} - \frac{1}{h_2} \overset{\circ}{y}_{x_2}.$$

Il est évident que

$$(R_1 + R_2) y = Ay = -\Lambda \overset{\bullet}{y} = -\overset{\circ}{y}_{\bar{x}_1 x_1} - \overset{\bullet}{y}_{\bar{x}_2 x_2}.$$

Le fait que les opérateurs R_1 et R_2 sont autoadjoints s'établit sans peine à l'aide de la formule de différences de Green. Comme il a été noté plus haut, dans le schéma (9) à la méthode de Seidel correspond la valeur $\tau = 2$, tandis qu'à la méthode de surrelaxation la valeur $\tau = 2/\sqrt{\delta(2-\delta)}$, où δ est défini dans (17).

De (11), (14) et du lemme 3 il s'ensuit que pour obtenir les estimations de la vitesse de convergence de ces méthodes dans H_A, il faut trouver δ et Δ à partir des inégalités (15). La constante δ est déjà trouvée. Cherchons Δ. A partir de la définition des opérateurs $\mathscr{D}$, R_1 et R_2, on obtient

$$(R_1 \mathscr{D}^{-1} R_2 y, y) = 0{,}5 \frac{h_1^2 h_2^2}{h_1^2 + h_2^2} (R_2 y, R_2 y). \tag{18}$$

Ensuite,

$$(R_2y,\ R_2y)=\frac{1}{h_1^2}(\mathring{y}_{\bar{x}_1}^2,\ 1)-\frac{2}{h_1h_2}(\mathring{y}_{\bar{x}_1},\ \mathring{y}_{\bar{x}_2})+\frac{1}{h_2^2}(\mathring{y}_{\bar{x}_2}^2,\ 1)\leqslant$$

$$\leqslant\left(\frac{1}{h_1^2}+\frac{1}{h_2^2}\right)[(\mathring{y}_{\bar{x}_1}^2,\ 1)+(\mathring{y}_{\bar{x}_2}^2,\ 1)]\leqslant\frac{h_1^2+h_2^2}{h_1^2h_2^2}(Ay,\ y).$$

En portant cette estimation dans (18), il vient

$$(R_1\mathcal{D}^{-1}R_2y,\ y)\leqslant\frac{1}{2}(Ay,\ y)$$

et, par conséquent, dans l'inégalité (15) $\Delta=2$.

Apprécions à présent la vitesse de convergence de la méthode de Seidel et de la méthode de surrelaxation.

De (11) il vient

$$\|z_n\|_A\leqslant\|S\|_A^n\|z_0\|_A$$

et, par suite, pour aboutir à la précision ε, il suffit d'effectuer $n\geqslant n_0(\varepsilon)$ itérations, où $n_0(\varepsilon)=\ln\varepsilon/\ln\|S\|_A$. De (14) on tire

$$n_0(\varepsilon)=2\ln\varepsilon/\ln\|S\|_A^2\geqslant\ln\frac{1}{\varepsilon}\Big/(\tau\gamma_1).\tag{19}$$

Pour la méthode de Seidel, de (16) on obtient ($\tau=2$)

$$\tau\gamma_1=2\delta/(1+4\delta)\tag{20}$$

et dans le cas particulier, quand $N_1=N_2=N$, $l_1=l_2=l$, on obtiendra de (17), (19) et (20)

$$\delta=2\sin^2\frac{\pi}{2N},\quad \tau\gamma_1\approx4\sin^2\frac{\pi}{2N}\approx\frac{\pi^2}{N^2},$$

$$n_0(\varepsilon)\approx\frac{N^2}{\pi^2}\ln\frac{1}{\varepsilon}\approx0{,}1N^2\ln\frac{1}{\varepsilon}.$$

Au point 1, § 5, ch. VI, on a obtenu pour la méthode explicite itérative simple appliquée à un cas particulier, l'estimation suivante du nombre d'itérations: $n_0(\varepsilon)\approx0{,}2N^2\ln\frac{1}{\varepsilon}$. En comparant ces estimations, on aboutit à ce que la méthode de Seidel exige environ deux fois moins d'itérations que la méthode itérative simple. Le caractère de la dépendance du nombre d'itérations de celui de nœuds N suivant une direction est le même pour ces deux méthodes, le nombre d'itérations est proportionnel à N^2.

Considérons maintenant la méthode de surrelaxation. En portant dans (16)

$$\delta=2\sin^2\frac{\pi}{2N},\quad \Delta=2\ \text{et}\ \tau=\frac{2}{\sqrt{\delta(2-\delta)}}=\frac{2}{\sin\frac{\pi}{N}},$$

on obtient

$$\tau\gamma_1 = \frac{2\,\mathrm{tg}\,\frac{\pi}{2N}}{2+2\,\mathrm{tg}\,\frac{\pi}{2N}+\mathrm{tg}^2\,\frac{\pi}{2N}} \approx \mathrm{tg}\,\frac{\pi}{2N} \approx \frac{\pi}{2N}.$$

De (19) tirons l'estimation suivante du nombre d'itérations pour la méthode de surrelaxation :

$$n_0(\varepsilon) \approx \frac{2N}{\pi}\ln\frac{1}{\varepsilon} \approx 0{,}64N\ln\frac{1}{\varepsilon}, \tag{21}$$

autrement dit, le nombre d'itérations pour la méthode de surrelaxation est proportionnel au nombre de nœuds N suivant une direction.

Pour conclure, donnons l'estimation du nombre d'itérations découlant du théorème 6. Pour la valeur du paramètre τ

$$\tau = \tau_0 = \frac{2}{\sqrt{\delta\Delta}} = \frac{1}{\sin\frac{\pi}{2N}}$$

le nombre d'itérations se définira par l'estimation

$$n \geqslant n_0(\varepsilon) = \ln\frac{1}{\varepsilon}\Big/\sin\frac{\pi}{2N} \approx 0{,}64N\ln\frac{1}{\varepsilon}.$$

Notons que l'estimation (21) est quelque peu surestimée. Pour s'en convaincre, il faut comparer les valeurs de $n_0(\varepsilon)$, calculées suivant la formule (21), avec le nombre d'itérations fourni au point 4 du § 2. La raison en est dans le fait que dans ce dernier cas le nombre d'itérations est apprécié sur la base de l'inégalité $\| S \|_A^n \leqslant \varepsilon$, tandis qu'au point 4 du § 2 les itérations se poursuivaient jusqu'à l'accomplissement de la condition $\| S^n \|_A \leqslant \varepsilon$.

CHAPITRE X

MÉTHODE TRIANGULAIRE ALTERNÉE

Ce chapitre est consacré à l'étude de la méthode itérative de la classe triangulaire alternée *) dans son application à la recherche de la solution de l'équation opératorielle à opérateur autoadjoint. On expose dans le § 1 la théorie générale de la méthode, on y décrit la construction du schéma itératif et l'on y indique le jeu des paramètres d'itération. La méthode est illustrée par un exemple du problème discret de Dirichlet pour l'équation de Poisson dans un rectangle. Au § 2 on montre comment cette méthode s'applique à la résolution des équations aux différences elliptiques à coefficients variables et à dérivées mixtes dans un rectangle. Dans le § 3 on a construit une variante de la méthode triangulaire alternée permettant de résoudre l'équation elliptique à coefficients variables sur un maillage irrégulier donné dans un domaine arbitraire.

§ 1. Théorie générale de la méthode

1. Schéma itératif. Au § 3 du ch. IX on a étudié la méthode itérative triangulaire de résolution de l'équation

$$Au = f. \tag{1}$$

Le schéma itératif de cette méthode est de la forme

$$B\,\frac{y_{k+1}-y_k}{\tau_{k+1}}+Ay_k=f,\quad k=0,\ 1,\ \ldots,\quad y_0\in H, \tag{2}$$

où $\tau_k \equiv \tau$, tandis que l'opérateur $B = B_1 = \mathscr{D} + \tau R_1$ se détermine par le développement suivant de l'opérateur A en une somme d'opérateurs

$$A = R_1 + R_2,\quad R_1 = R_2^*,\quad A = A^* > 0. \tag{3}$$

En ce qui concerne l'opérateur $\mathscr{D}$, on suppose qu'il est autoadjoint et défini positif dans H, c'est-à-dire que

$$\mathscr{D} = \mathscr{D}^* > 0. \tag{4}$$

La méthode itérative triangulaire appartient à la classe des méthodes dont les paramètres d'itération sont choisis compte tenu de l'information à priori sur les opérateurs du schéma itératif. Pour la méthode triangulaire, l'information initiale comporte la fixation des

*) La méthode a été proposée par A. A. Samarski en 1964 (voir ЖВМ et МФ, 4, n° 3, 1964) et améliorée dans [8].

constantes δ et Δ des inégalités

$$\delta\mathscr{D}\leqslant A, \quad R_1\mathscr{D}^{-1}R_2\leqslant\frac{\Delta}{4}A, \quad \delta>0. \tag{5}$$

Le paramètre τ trouvé au point 3, § 3, ch. IX, permet d'atteindre la précision ε en $n_0 = O(\ln(1/\varepsilon)\sqrt{\eta})$ itérations, où $\eta = \delta/\Delta$.

Notons que le fait que l'opérateur B n'est pas autoadjoint ne permet pas d'utiliser dans le schéma itératif (2) le jeu des paramètres τ_k et, partant, d'augmenter la vitesse de convergence de la méthode. Cependant la simplicité de la construction de l'opérateur B et la possibilité du développement (3) pour l'opérateur A, quelle que soit sa structure, ont constitué une stimulation à l'étude des variantes possibles de la méthode triangulaire. On a ainsi fini par construire la méthode triangulaire alternée cumulant l'universalité de la construction de l'opérateur B avec la possibilité de choix dans le schéma (2) du jeu des paramètres τ_k.

Abordons donc l'étude de la méthode triangulaire alternée. Le schéma itératif de la méthode a l'aspect (2), où l'opérateur B se définit de la façon suivante:

$$B = (\mathscr{D} + \omega R_1)\mathscr{D}^{-1}(\mathscr{D} + \omega R_2), \quad \omega > 0. \tag{6}$$

ω est ici le paramètre d'itération qu'on doit définir. Supposons ensuite, que pour le schéma (2), (6) les conditions (3), (4) sont remplies et que δ et Δ sont fixés dans les inégalités (5).

Notons quelques propriétés de l'opérateur B défini par la relation (6). Si à l'opérateur $\mathscr{D} + \omega R_1$ correspond une matrice triangulaire et à $\mathscr{D}$ une matrice diagonale, à B correspond alors le produit de deux matrices triangulaires et une matrice diagonale. Dans ce cas l'inversion de l'opérateur B n'est pas une opération laborieuse.

Montrons que l'opérateur B est autoadjoint dans H, et si l'opérateur $\mathscr{D}$ est borné, B est défini positif. En effet, en vertu de (3) on a l'égalité

$$(Au, u) = 2(R_1u, u) = 2(R_2u, u) > 0.$$

De là et à partir de (4) il s'ensuit que les opérateurs $B_1 = \mathscr{D} + \omega R_1$ et $B_2 = \mathscr{D} + \omega R_2$ sont autoadjoints et définis positifs: $B_1^* = (\mathscr{D} + \omega R_1)^* = \mathscr{D} + \omega R_2 = B_2$, $B_\alpha > \mathscr{D} > 0$, $\alpha = 1, 2$; et, par suite,

$$B^* = (B_1\mathscr{D}^{-1}B_2)^* = B_2^*\mathscr{D}^{-1}B_1^* = B_1\mathscr{D}^{-1}B_2 = B.$$

Ensuite, puisque $\mathscr{D}$ est un opérateur autoadjoint, borné et défini positif, l'opérateur inverse $\mathscr{D}^{-1}$ sera défini positif dans H. Par conséquent, en utilisant l'inégalité $(\mathscr{D}^{-1}x, x)\geqslant d(x, x)$, $d>0$ signifiant que l'opérateur $\mathscr{D}^{-1}$ est défini positif, il vient

$$(Bu, u) = (\mathscr{D}^{-1}B_2u, B_2u)\geqslant d\,\|B_2u\|^2 > 0.$$

Il découle de (2), (6) que pour définir y_{k+1}, y_k une fois donné, il faut résoudre l'équation

$$(\mathscr{D}+\omega R_1)\mathscr{D}^{-1}(\mathscr{D}+\omega R_2)y_{k+1}=\varphi_k, \quad k=0, 1, \ldots,$$

où $\varphi_k = By_k - \tau_{k+1}(Ay_k - f)$. Cela se réduit à la résolution de deux équations

$$(\mathscr{D}+\omega R_1)\,v=\varphi_k, \quad (\mathscr{D}+\omega R_2)\,y_{k+1}=\mathscr{D}v.$$

Pour la mise en œuvre du schéma (2), (6), on peut faire appel à un second algorithme basé sur l'écriture du schéma sous forme d'un schéma avec correction

$$y_{k+1}=y_k-\tau_{k+1}w_k, \quad Bw_k=r_k,$$

où $r_k = Ay_k - f$ est le résidu. La correction w_k est recherchée par résolution de deux équations

$$(\mathscr{D}+\omega R_1)\,\bar{w}_k=r_k, \quad (\mathscr{D}+\omega R_2)\,w_k=\mathscr{D}\bar{w}_k.$$

Dans cet algorithme on s'est dispensé de calculer By_k, mais, par contre, on est obligé de mémoriser simultanément y_k et les grandeurs intermédiaires r_k, $\bar{w}_k$, w_k.

2. Choix des paramètres d'itération. Passons à présent à l'étude de la convergence du schéma itératif (2), (6). Vu que les opérateurs A et B sont autoadjoints et définis positifs dans H, on est en mesure d'étudier la convergence dans H_D, où en guise de D on a pris l'un des opérateurs A, B ou $AB^{-1}A$ (dans ce dernier cas B doit être un opérateur borné). Pour l'opérateur D mentionné l'opérateur $DB^{-1}A$ sera apparemment autoadjoint dans H et, par suite, selon la classification établie au ch. VI, on a un schéma itératif avec opérateur autoadjoint.

En profitant des résultats obtenus au § 2 du ch. VI, on peut aussitôt indiquer pour le schéma (2), (6) le jeu optimal de paramètres d'itération τ_k. Soient γ_1 et γ_2 empruntés aux inégalités

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \quad \gamma_1 > 0. \tag{7}$$

Dans ce cas le jeu des paramètres de Tchébychev $\{\tau_k\}$ se détermine suivant les formules

$$\begin{gathered}\tau_k=\frac{\tau_0}{1+\rho_0\mu_k}, \quad \mu_k\in\mathfrak{M}_n^*=\left\{\cos\frac{(2i-1)\,\pi}{2n}, \ 1\leqslant i\leqslant n\right\}, \ 1\leqslant k\leqslant n,\\ \tau_0=\frac{2}{\gamma_1+\gamma_2}, \quad \rho_0=\frac{1-\xi}{1+\xi}, \quad n\geqslant n_0(\varepsilon)=\frac{\ln(0{,}5\varepsilon)}{\ln\rho_1}, \quad \xi=\frac{\gamma_1}{\gamma_2},\end{gathered} \tag{8}$$

et pour l'erreur $z_n = y_n - u$ de la méthode itérative (2), (6), (8) on a l'estimation

$$\|z_n\|_D\leqslant q_n\|z_0\|_D, \quad q_n=\frac{2\rho_1^n}{1+\rho_1^{2n}}\leqslant\varepsilon, \quad \rho_1=\frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}. \tag{9}$$

C'est le résultat de la théorie générale des méthodes itératives à deux couches. Pour le schéma (2), (6) l'information à priori est fournie par les constantes δ et Δ des inégalités (5). Aussi un des problèmes est l'obtention des expressions de γ_1 et γ_2 au moyen de δ et Δ. Ensuite, puisque l'opérateur B dépend du paramètre d'itération ω, γ_1 et γ_2 sont des fonctions de $\omega : \gamma_1 = \gamma_1(\omega)$, $\gamma_2 = \gamma_2(\omega)$. Vu que de l'estimation (9) il s'ensuit que la vitesse maximale de convergence sera atteinte quand le rapport $\xi = \gamma_1/\gamma_2$ est maximal, on aboutit au problème du choix du paramètre ω sur la base de la condition du maximum de ξ. Les problèmes formulés sont résolus avec le lemme 1.

L e m m e 1. *Soient remplies les conditions* (3), (4), *l'opérateur B étant défini par la formule* (6), *et dans les inégalités* (5) *les constantes δ et Δ sont fixées. On a alors dans les inégalités* (7)

$$\gamma_1 = \delta\Big/\left(1 + \omega\delta + \frac{1}{4}\,\omega^2\delta\Delta\right), \quad \gamma_2 = 1/(2\omega). \tag{10}$$

Le rapport $\xi = \gamma_1/\gamma_2$ est maximal si

$$\omega = \omega_0 = 2/\sqrt{\delta\Delta}, \tag{11}$$

avec

$$\gamma_1 = \frac{\delta}{2(1+\sqrt{\eta})}, \quad \gamma_2 = \frac{\delta}{4\sqrt{\eta}}, \quad \xi = \frac{2\sqrt{\eta}}{1+\sqrt{\eta}}, \quad \eta = \frac{\delta}{\Delta}. \tag{12}$$

Et de fait, écrivons l'opérateur B sous forme

$$B = (\mathscr{D} + \omega R_1)\,\mathscr{D}^{-1}(\mathscr{D} + \omega R_2) = \mathscr{D} + \omega(R_1 + R_2) + \\ + \omega^2 R_1\mathscr{D}^{-1}R_2. \tag{13}$$

Compte tenu de ce que $A = R_1 + R_2 \geqslant \delta\mathscr{D}$ ou $\mathscr{D} \leqslant \frac{1}{\delta} A$, on obtient pour B l'estimation par le haut

$$B \leqslant \left(\frac{1}{\delta} + \omega + \frac{1}{4}\,\omega^2\Delta\right) A = \frac{1}{\gamma_1} A,$$

c'est-à-dire $A \geqslant \gamma_1 B$, où γ_1 est défini dans (10).

Transformons maintenant la formule (13):

$$B = \mathscr{D} - \omega(R_1 + R_2) + \omega^2 R_1\mathscr{D}^{-1}R_2 + 2\omega(R_1 + R_2) = \\ = (\mathscr{D} - \omega R_1)\,\mathscr{D}^{-1}(\mathscr{D} - \omega R_2) + 2\omega A.$$

Il s'ensuit

$$(By, y) = 2\omega(Ay, y) + (\mathscr{D}^{-1}(\mathscr{D} - \omega R_2)\,y, (\mathscr{D} - \omega R_2)\,y).$$

En utilisant le fait que l'opérateur $\mathscr{D}^{-1}$ est défini positif, on obtient $(By, y) \geqslant 2\omega\,(Ay, y)$, c'est-à-dire $A \leqslant \gamma_2 B$. Bref, on a obtenu γ_1 et γ_2. Examinons ensuite la relation

$$\xi = \xi(\omega) = \gamma_1/\gamma_2 = 2\omega\delta\Big/\left(1 + \omega\delta + \frac{\omega^2\delta\Delta}{4}\right).$$

En annulant la dérivée

$$\xi'(\omega)=\frac{2\delta(1-\omega^2\delta\Delta/4)}{\left(1+\omega\delta+\frac{\omega^2\delta\Delta}{4}\right)^2},$$

on trouve $\omega=\omega_0=2/\sqrt{\delta\Delta}$. En ce point $\xi(\omega)$ devient maximum, car $\xi''(\omega_0)<0$. En portant ω trouvé dans (10), on aboutit à (12). Montrons que $\delta\leqslant\Delta$, $\eta\leqslant 1$. En effet, en utilisant l'égalité $(Ax, x)=2(R_2x, x)$ et l'inégalité de Cauchy-Bouniakovski, on obtient de (5)

$$\delta(\mathscr{D}x, x)\leqslant(Ax, x)=\frac{(Ax, x)^2}{(Ax, x)}=4\frac{(R_2x, x)^2}{(Ax, x)}=$$

$$=4\frac{(\mathscr{D}^{-1/2}R_2x, \mathscr{D}^{1/2}x)^2}{(Ax, x)}\leqslant 4\frac{(\mathscr{D}^{-1/2}R_2x, \mathscr{D}^{-1/2}R_2x)}{(Ax, x)}(\mathscr{D}^{1/2}x, \mathscr{D}^{1/2}x)=$$

$$=4\frac{(R_1\mathscr{D}^{-1}R_2x, x)}{(Ax, x)}(\mathscr{D}x, x)\leqslant\Delta(\mathscr{D}x, x),$$

ce qu'il fallait démontrer. Le lemme est démontré.

Théorème 1. *Soient remplies les conditions du lemme* 1. *Alors pour la méthode triangulaire alternée* (2), (6), (11) *avec les paramètres de Tchébychev* τ_k *définis par les formules* (8) *et* (12) *se vérifie l'estimation* (9). *Pour que l'inégalité* $\|z_n\|_D\leqslant\varepsilon\|z_0\|_D$ *soit satisfaite il suffit d'effectuer n itérations, où* $n\geqslant n_0(\varepsilon)$, $n_0(\varepsilon)=\ln\frac{2}{\varepsilon}\Big/\left(2\sqrt{2}\sqrt[4]{\eta}\right)$ $\eta=\delta/\Delta$. *On a dans ce cas* $D=A$, B *ou* $AB^{-1}A$.

Pour démontrer le théorème, il faut utiliser le lemme 1 et les formules (8) fournissant les paramètres d'itération et les nombres d'itérations.

Examinons à présent un procédé utilisé pour la construction des schémas itératifs implicites. Soit donné dans H l'opérateur R autoadjoint et défini positif qui est énergétiquement équivalent à l'opérateur A à constantes c_1 et c_2:

$$c_1R\leqslant A\leqslant c_2R,\quad c_1>0, \tag{14}$$

et à l'opérateur B à constantes $\mathring{\gamma}_1$ et $\mathring{\gamma}_2$:

$$\mathring{\gamma}_1B\leqslant R\leqslant\mathring{\gamma}_2B,\quad \mathring{\gamma}_1>0. \tag{15}$$

Supposons que les opérateurs A et B sont autoadjoints. On obtient pour ces derniers de (14) et (15) les inégalités suivantes: $\gamma_1B\leqslant A\leqslant\gamma_2B$, $\gamma_1=c_1\mathring{\gamma}_1$, $\gamma_2=c_2\mathring{\gamma}_2$. Le procédé exposé permet, lors de la construction de l'opérateur B, de partir non pas du développement (3) de l'opérateur A, mais du développement de l'opérateur R, qui peut être choisi le même pour une grande classe d'opérateurs A variés. De plus, les constantes $\mathring{\gamma}_1$ et $\mathring{\gamma}_2$ peuvent être choisies dans (15)

une seule fois. Le problème d'acquisition pour la méthode de l'information à priori se réduit ainsi à la recherche de c_1 et c_2 dans (14).

Bref, soit l'opérateur R représenté sous forme de somme d'opérateurs autoadjoints R_1 et R_2:

$$R = R^* > 0, \quad R = R_1 + R_2, \quad R_1 = R_2^*. \tag{16}$$

et au lieu de (5) on a les inégalités

$$\delta \mathcal{D} \leqslant R, \quad R_1 \mathcal{D}^{-1} R_2 \leqslant \frac{\Delta}{4} R, \quad \delta > 0. \tag{17}$$

L'opérateur B pour le schéma (2) sera construit suivant la formule (6). Alors, en vertu du lemme 1, pour $\omega = \omega_0 = 2/\sqrt{\delta\Delta}$ dans les inégalités (15) on a

$$\overset{\circ}{\gamma}_1 = \frac{\delta}{2(1+\sqrt{\eta})}, \quad \overset{\circ}{\gamma}_2 = \frac{\delta}{4\sqrt{\eta}}, \quad \overset{\circ}{\xi} = \frac{\overset{\circ}{\gamma}_1}{\overset{\circ}{\gamma}_2} = \frac{2\sqrt{\eta}}{1+\sqrt{\eta}}, \quad \eta = \frac{\delta}{\Delta}. \tag{18}$$

Il s'ensuit de là le

Théorème 2. *Soient* $A = A^* > 0$, $\mathcal{D} = \mathcal{D}^* > 0$, *les conditions* (16) *étant remplies*, c_1 *et* c_2 *donnés dans* (14) *et* δ, Δ *dans* (17). *Alors l'estimation* (9) *est satisfaite pour la méthode triangulaire alternée* (2), (6), (8), (11) *avec les paramètres de Tchébychev* τ_k, *où* $\gamma_1 = c_1\overset{\circ}{\gamma}_1$ *et* $\gamma_2 = c_2\overset{\circ}{\gamma}_2$, $\overset{\circ}{\gamma}_1$ *et* $\overset{\circ}{\gamma}_2$ *étant définis dans* (18). *Pour la satisfaction de l'inégalité* $\| z_n \|_D \leqslant \varepsilon \| z_0 \|_D$ *on peut se contenter de* n *itérations avec*

$$n \geqslant n_0(\varepsilon), \quad n_0(\varepsilon) = \frac{\ln(2/\varepsilon)}{2\sqrt{2}\sqrt[4]{\eta}} \sqrt{\frac{c_2}{c_1}}, \quad \eta = \frac{\delta}{\Delta}.$$

3. Méthode d'obtention des grandeurs initiales δ et Δ. Des théorèmes 1 et 2 il s'ensuit que pour l'application de la méthode triangulaire alternée il faut fixer deux nombres δ et Δ dans les inégalités (5) ou (17). Dans les exemples fournis plus bas des équations de mailles elliptiques ces constantes seront obtenues sous la forme explicite ou on donnera les algorithmes permettant de les calculer. Il va de soi que dans ce cas on utilisera la structure des opérateurs A, R_1, R_2 et $\mathcal{D}$. Pour la théorie générale des méthodes itératives, qui ne tient pas compte de la structure des opérateurs, il est nécessaire de fournir un procédé général d'obtention de l'information à priori exigée pour la mise en œuvre de la méthode.

Ce procédé peut se baser sur l'utilisation de la propriété asymptotique des méthodes itératives du type variationnel (voir point 5, § 1, ch. VIII). Supposons que les opérateurs A et B sont autoadjoints et définis positifs dans H. Si dans le schéma itératif

$$B\frac{v_{k+1} - v_k}{\tau_{k+1}} + Av_k = 0, \quad k = 0, 1, \ldots, \quad v_0 \neq 0 \tag{19}$$

on choisit les paramètres τ_{k+1} suivant la formule de la méthode de la plus grande pente

$$\tau_{k+1} = \frac{(w_k, r_k)}{(Aw_k, w_k)}, \quad k = 0, 1, \ldots, \quad r_k = Av_k, \quad Bw_k = r_k, \qquad (20)$$

et pour un numéro d'itérations n suffisamment grand, on recherche les racines $x_1 \leqslant x_2$ de l'équation

$$(1 - \tau_n x)(1 - \tau_{n-1} x) = \rho_n \rho_{n-1}, \quad \rho_n = \frac{\| v_n \|_A}{\| v_{n-1} \|_A}, \qquad (21)$$

où $\| \cdot \|_A$ est la norme dans H_A, alors x_1 et x_2 seront des approximations de γ_1 et γ_2 dans les inégalités (7) respectivement par le haut et par le bas.

Profitons du procédé décrit. Considérons le schéma itératif (2), (3), (6). Notons qu'en vertu du lemme 1 dans les inégalités (7) $\gamma_2 = 1/(2\omega)$, tandis que de l'information à priori ne dépend que γ_1. On tâchera, sans chercher séparément δ et Δ des inégalités (5), de trouver d'emblée l'expression de γ_1 comme une fonction du paramètre d'itération ω. On donnera à cette expression la forme (10) en indiquant les valeurs correspondantes de δ et Δ. Alors, sur la base du lemme 1, on obtient ω_0 au moyen de la formule (11), ainsi que γ_1 et γ_2 correspondant à ω_0 suivant la formule (12). Pour avoir le jeu des paramètres τ_k on utilise (8).

Obtenons l'expression cherchée de γ_1. Posons $\omega = 0$ et, suivant la méthode (19)-(21), cherchons x_1. En supposant que dans (19), (20) on a effectué un nombre suffisant d'itérations et compte tenu de ce que l'opérateur $B = \mathscr{D}$ pour $\omega = 0$, on obtient l'inégalité approchée

$$x_1 \mathscr{D} \leqslant A, \quad x_1 > 0. \qquad (22)$$

Ensuite, posons $\omega = \omega_1 > 0$ et, suivant la méthode (19)-(21), cherchons $\bar{x}_1$, de manière que $\bar{x}_1 > 0$ et

$$\bar{x}_1 B \leqslant A \quad \text{ou} \quad \bar{x}_1 (\mathscr{D} + \omega_1 A + \omega_1^2 R_1 \mathscr{D}^{-1} R_2) \leqslant A, \qquad (23)$$

en outre, on voit que $\bar{x}_1 \omega_1 < 1$. Ecrivons (23) sous la forme

$$\bar{x}_1 \mathscr{D} + \bar{x}_1 \omega_1^2 R_1 \mathscr{D}^{-1} R_2 \leqslant (1 - \bar{x}_1 \omega_1) A$$

et additionnons-la à (22) qui, au préalable, est multipliée par un coefficient $\alpha > 0$ indéterminé pour le moment. On obtient

$$(\alpha x_1 + \bar{x}_1) \mathscr{D} + \bar{x}_1 \omega_1^2 R_1 \mathscr{D}^{-1} R_2 \leqslant (1 - \bar{x}_1 \omega_1 + \alpha) A. \qquad (24)$$

Divisons cette inégalité par $\alpha x_1 + \bar{x}_1$, ajoutons aux second et premier membres le terme ωA et choisissons α de la condition

$$\bar{x}_1 \omega_1^2 = \omega^2 (\alpha x_1 + \bar{x}_1); \qquad (25)$$

alors l'inégalité transformée prendra la forme

$$\mathscr{D} + \omega A + \omega^2 R_1 \mathscr{D}^{-1} R_2 = B \leqslant \frac{1}{\gamma_1} A,$$

où

$$\frac{1}{\gamma_1}=\frac{1}{\gamma_1(\omega)}=\omega+\frac{1-\bar{x}_1\omega_1+\alpha}{\alpha x_1+\bar{x}_1}. \qquad (26)$$

De (25) on tire α:

$$\alpha=\bar{x}_1(\omega_1^2-\omega^2)/(\omega^2 x_1).$$

Vu que α doit être positif, l'expression (26) n'aura lieu que pour $0<\omega<\omega_1$. En portant α trouvé dans (26), il vient

$$\frac{1}{\gamma_1}=\frac{1}{x_1}+\omega+\frac{x_1-\bar{x}_1-x_1\bar{x}_1\omega_1}{x_1\bar{x}_1\omega_1^2}\,\omega^2.$$

En comparant cette expression à (10), on constate qu'en guise de δ et Δ on peut prendre

$$\delta=x_1, \quad \Delta=4\frac{x_1-\bar{x}_1-x_1\bar{x}_1\omega_1}{x_1\bar{x}_1\omega_1^2}.$$

Notons que ω_0, obtenu avec les δ et Δ mentionnés selon (11), sera compris dans l'intervalle $(0, \omega_1)$ si l'inégalité $2\bar{x}_1\leqslant x_1(1-\bar{x}_1\omega_1)$ est satisfaite. Si cette inégalité n'est pas satisfaite il faut augmenter ω_1 et reprendre les calculs indiqués (on recommande de prendre $\omega_1=2/x_1$).

4. Problème discret de Dirichlet pour l'équation de Poisson dans un rectangle. Illustrons par l'exemple d'un problème discret de Dirichlet pour l'équation de Poisson dans le rectangle $\bar{G}=\{0\leqslant x_\alpha\leqslant l_\alpha, \alpha=1, 2\}$ l'application de la méthode triangulaire alternée

$$\begin{aligned}\Lambda y=y_{\bar{x}_1x_1}+y_{\bar{x}_2x_2}&=-\varphi(x), \quad x\in\omega,\\ y(x)&=g(x), \quad x\in\gamma\end{aligned} \qquad (27)$$

sur le maillage $\bar{\omega}=\{x_{ij}=(ih_1, jh_2)\in\bar{G}, 0\leqslant i\leqslant N_1, 0\leqslant j\leqslant N_2, h_\alpha=l_\alpha/N_\alpha, \alpha=1, 2\}$ à frontière γ.

Pour l'exemple considéré H est un espace des fonctions de mailles associées à ω avec produit scalaire

$$(u, v)=\sum_{x\in\omega}u(x)v(x)h_1h_2.$$

L'opérateur A se détermine par l'égalité $Ay=-\Lambda\dot{y}$, où $y\in H$, $\mathring{y}\in\mathring{H}$ et $y(x)=\mathring{y}(x)$, $x\in\omega$, tandis que $\mathring{y}(x)=0$ pour $x\in\gamma$. Le

second membre f sera déterminé de façon ordinaire: $f(x) = \varphi(x) + \frac{1}{h_1^2}\varphi_1(x) + \frac{1}{h_2^2}\varphi_2(x)$, où

$$\varphi_1(x) = \begin{cases} g(0, x_2), & x_1 = h_1, \\ 0, & 2h \leqslant x_1 \leqslant l_1 - 2h_1, \\ g(l_1, x_2), & x_1 = l_1 - h_1, \end{cases}$$

$$\varphi_2(x) = \begin{cases} g(x_1, 0), & x_2 = h_2, \\ 0, & 2h_2 \leqslant x_2 \leqslant l_2 - 2h_2, \\ g(x_1, l_2), & x_2 = l_2 - h_2. \end{cases}$$

Dans ce cas le problème (27) s'écrit sous la forme de l'équation (1).

L'opérateur A est autoadjoint et défini positif dans H, car il correspond à l'opérateur de différences de Laplace avec les conditions aux limites de Dirichlet.

Passons maintenant à la construction de l'opérateur B. Soit la variante classique de la méthode triangulaire alternée dans laquelle on pose dans (6)

$$\mathscr{D} = E. \tag{28}$$

Déterminons à présent les opérateurs de différences $\mathscr{R}_1$ et $\mathscr{R}_2$ qui agissent sur les fonctions de mailles associées à $\bar{\omega}$ de la façon suivante:

$$\mathscr{R}_1 y = -\sum_{\alpha=1}^{2} \frac{1}{h_\alpha} y_{\bar{x}_\alpha}, \quad \mathscr{R}_2 y = \sum_{\alpha=1}^{2} \frac{1}{h_\alpha} y_{x_\alpha}, \quad x \in \omega.$$

Apparemment, $\mathscr{R}_1 + \mathscr{R}_2 = \Lambda$. En utilisant les formules aux différences de Green, on constate sans peine que pour les fonctions de mailles $\mathring{y}(x) \in \mathring{H}$, $\mathring{u}(x) \in \mathring{H}$, c'est-à-dire associées à $\bar{\omega}$ et s'annulant sur γ, on a l'égalité

$$(\mathscr{R}_1 \mathring{y}, \mathring{u}) = (\mathring{y}, \mathscr{R}_2 \mathring{u}). \tag{29}$$

Définissons sur H les opérateurs R_1 et R_2 de la façon suivante: $R_\alpha y = -\mathscr{R}_\alpha \mathring{y}$, $\alpha = 1, 2$, où $y \in H$, $\mathring{y} \in \mathring{H}$ et $y(x) = \mathring{y}(x)$, $x \in \omega$. Dans ce cas, en vertu de la définition des opérateurs de différences $\mathscr{R}_\alpha$ et de l'égalité (29), les conditions (3) sont remplies, c'est-à-dire que $A = R_1 + R_2$, $R_1 = R_2^*$. Compte tenu de (28), on obtient de (6) la forme suivante de l'opérateur B:

$$B = (E + \omega R_1)(E + \omega R_2).$$

Cherchons à présent l'information à priori nécessaire pour la mise en œuvre de la méthode triangulaire alternée. Dans le cas considéré elle a la forme des constantes δ et Δ des inégalités $\delta E \leqslant A$, $R_1 R_2 \leqslant (\Delta/4)A$. Apparemment on peut prendre en guise de δ la

valeur propre minimale de l'opérateur de différences de Laplace

$$\delta=\frac{4}{h_1^2}\sin^2\frac{\pi h_1}{2l_1}+\frac{4}{h_2^2}\sin^2\frac{\pi h_2}{2l_2}.$$

L'estimation de Δ a été trouvée au point 4, § 3, ch. IX (les opérateurs R_1 et R_2, définis dans les deux cas, coïncident). On a $\Delta = 4/h_1^2 + 4/h_2^2$.

En résumé, l'information sur δ et Δ est obtenue. Du lemme 1 déduisons la valeur optimale du paramètre ω_0, ainsi que γ_1 et γ_2. Les paramètres d'itération τ_k se calculent suivant les formules (8). Dans le cas particulier, où $N_1 = N_2 = N$, $l_1 = l_2 = l$, il vient

$$\delta=\frac{8}{h^2}\sin^2\frac{\pi}{2N},\quad \Delta=\frac{8}{h^2},\quad \eta=\frac{\delta}{\Delta}=\sin^2\frac{\pi}{2N},$$

$$\xi=\frac{2\sqrt{\eta}}{1+\sqrt{\eta}}\approx 2\sqrt{\eta}=2\sin\frac{\pi}{2N}\approx\frac{\pi}{N},\qquad \omega_0=\frac{h^2}{4\sin\frac{\pi h}{2l}}.$$

Suivant le théorème 1, on aura dans le cas considéré pour le nombre d'itérations $n\geqslant n_0(\varepsilon)$, où

$$n_0(\varepsilon)=\frac{\ln(2/\varepsilon)}{2\sqrt{2}\sqrt[4]{\eta}}\approx\frac{\sqrt{N}}{2\sqrt{\pi}}\ln\frac{2}{\varepsilon}\approx 0{,}28\sqrt{N}\ln\frac{2}{\varepsilon},$$

c'est-à-dire que le nombre d'itérations est proportionnel à la racine d'indice quatre du nombre d'inconnues dans le problème.

Au point 4, § 3, ch. IX, on a obtenu pour la méthode de surrelaxation, appliquée à la résolution du problème de différences (27), l'estimation suivante du nombre d'itérations:

$$n\geqslant n_0(\varepsilon)\approx 0{,}64N\ln(1/\varepsilon),\qquad (N=l/h).$$

La comparaison de la méthode de relaxation avec la méthode triangulaire alternée fait ressortir l'avantage évident de cette dernière. Bien que la mise en œuvre d'une itération dans la méthode triangulaire alternée exige un nombre deux fois plus grand d'opérations arithmétiques que celle de la méthode de relaxation, cette méthode présente un avantage substantiel sous l'angle du nombre d'itérations ce qui garantit l'efficience générale de cette méthode.

Donnons maintenant le nombre d'itérations que coûte la méthode triangulaire alternée avec les paramètres de Tchébychev, appliquée au problème de différences (27), en fonction du nombre de nœuds N suivant une direction du maillage carré $\bar{\omega}$ pour $\varepsilon = 10^{-4}$:

$N = 32 \quad n = 16$
$N = 64 \quad n = 23$
$N = 128 \quad n = 32$

La comparaison avec le nombre d'itérations de la méthode de surrelaxation, effectuée au point 4, § 2, ch. IX, montre que la méthode

de relaxation exige environ 3,5-7,5 fois plus d'itérations que la méthode triangulaire alternée.

Remarque 1. Si au lieu d'un rectangle $\overline{G}$ on prend un parallélépipède à p dimensions $\overline{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha,\ \alpha = 1, 2, \ldots, p\}$ sur lequel est donné le problème discret de Dirichlet pour l'équation de Poisson

$$\Lambda y = \sum_{\alpha=1}^{p} y_{\bar{x}_\alpha x_\alpha} = -\varphi(x), \quad x \in \omega,$$

$$y(x) = g(x), \quad x \in \gamma,$$

associée au maillage rectangulaire $\bar{\omega} = \{x_i = (i_1 h_1, i_2 h_2, \ldots, i_p h_p) \in \in \overline{G},\ 0 \leqslant i_\alpha \leqslant N_\alpha,\ h_\alpha N_\alpha = l_\alpha,\ \alpha = 1, 2, \ldots, p\}$, alors les opérateurs de différences $\mathcal{R}_1$ et $\mathcal{R}_2$ se définiront ainsi:

$$\mathcal{R}_1 y = -\sum_{\alpha=1}^{p} \frac{1}{h_\alpha} y_{\bar{x}_\alpha}, \quad \mathcal{R}_2 y = \sum_{\alpha=1}^{p} \frac{1}{h_\alpha} y_{x_\alpha}, \quad x \in \omega.$$

Dans ce cas dans les inégalités (5) on doit poser pour $\mathcal{D} = E$

$$\delta = \sum_{\alpha=1}^{p} \frac{1}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2 l_\alpha}, \quad \Delta = \sum_{\alpha=1}^{p} \frac{4}{h_\alpha^2},$$

car

$$\| \mathcal{R}_2 \mathring{y} \|^2 = (\mathcal{R}_1 \mathcal{R}_2 \mathring{y}, \mathring{y}) = \left\| \sum_{\alpha=1}^{p} \frac{1}{h_\alpha} \mathring{y}_{x_\alpha} \right\| \leqslant$$

$$\leqslant \sum_{\alpha=1}^{p} \frac{1}{h_\alpha^2} \sum_{\alpha=1}^{p} \| \mathring{y}_{x_\alpha} \|^2 \leqslant \left(\sum_{\alpha=1}^{p} \frac{1}{h_\alpha^2} \right) (A\mathring{y}, \mathring{y}).$$

En outre, au cas de $N_1 = N_2 = \ldots = N_p = N$, $l_1 = l_2 = \ldots = l_p = l$, on a pour le nombre d'itérations l'estimation

$$n \geqslant n_0(\varepsilon), \quad n_0(\varepsilon) \approx \frac{\sqrt{N}}{2\sqrt{\pi}} \ln \frac{2}{\varepsilon} \approx 0{,}28 \sqrt{N} \ln \frac{2}{\varepsilon},$$

qui ne dépend pas du nombre de dimensions p.

Passons à présent aux questions se rapportant à la mise en œuvre de la méthode triangulaire alternée pour le problème (27). Au point 1 on a fourni deux algorithmes permettant de trouver y_{k+1} pour le schéma itératif de la méthode une fois donné y_k. Considérons d'abord le

second algorithme. Pour $\mathscr{D} = E$ il acquiert l'aspect

$$\begin{gathered} r_k = Ay_k - f, \\ (E + \omega_0 R_1)\,\overline{w}_k = r_k, \quad (E + \omega_0 R_2)\,w_k = \overline{w}_k, \qquad (30) \\ y_{k+1} = y_k - \tau_{k+1} w_k, \quad k = 0, 1, \ldots \end{gathered}$$

Cet algorithme doit être utilisé quand les paramètres τ_k sont choisis non pas d'après les formules (8), mais au moyen des formules des méthodes itératives du type variationnel.

En profitant de la définition des opérateurs A, R_1 et R_2 au moyen des opérateurs de différences Λ, $\mathscr{R}_1$ et $\mathscr{R}_2$, on peut récrire les formules (30) sous la forme suivante:

$$\begin{gathered} r_k(i, j) = \left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right) y_k(i, j) - \frac{1}{h_1^2}[y_k(i-1, j) + y_k(i+1, j)] - \\ - \frac{1}{h_2^2}[y_k(i, j-1) + y_k(i, j+1)] - \varphi(i, j), \qquad (31) \\ 1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant j \leqslant N_2 - 1, \quad y_k|_\gamma = g. \end{gathered}$$

$$\begin{gathered} \overline{w}_k(i, j) = \alpha \overline{w}_k(i-1, j) + \beta \overline{w}_k(i, j-1) + \varkappa r_k(i, j), \\ i = 1, 2, \ldots, N_1 - 1, \quad j = 1, 2, \ldots, N_2 - 1, \qquad (32) \\ \overline{w}_k(0, j) = 0, \quad 1 \leqslant j \leqslant N_2 - 1, \quad \overline{w}_k(i, 0) = 0, \ 1 \leqslant i \leqslant N_1 - 1. \end{gathered}$$

Le compte est dans ce cas mené à partir du point $i = 1$, $j = 1$, ou bien suivant les lignes du maillage $\overline{\omega}$, c'est-à-dire avec l'accroissement de i pour un j fixé ou suivant les colonnes avec l'accroissement de j pour un i fixé

$$\begin{gathered} w_k(i, j) = \alpha w_k(i+1, j) + \beta w_k(i, j+1) + \varkappa \overline{w}_k(i, j), \\ i = N_1 - 1, N_1 - 2, \ldots, 1, \quad j = N_2 - 1, N_2 - 2, \ldots, 1, \qquad (33) \\ w_k(N_1, j) = 0, \quad 1 \leqslant j \leqslant N_2 - 1, \quad w_k(i, N_2) = 0, \\ 1 \leqslant i \leqslant N_1 - 1. \end{gathered}$$

Ici le compte est mené à partir du point $i = N_1 - 1$, $j = N_2 - 1$, soit par lignes, soit par colonnes du maillage avec le décroissement de l'indice i ou j respectivement. Finalement, y_{k+1} s'obtient suivant les formules

$$\begin{gathered} y_{k+1}(i, j) = y_k(i, j) - \tau_{k+1} w_k(i, j), \\ 1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant j \leqslant N_2 - 1, \qquad (34) \\ y_{k+1}|_\gamma = g. \end{gathered}$$

On a utilisé dans ce cas les notations suivantes:

$$\alpha = \frac{\omega_0 h_2^2}{h_1^2 h_2^2 + \omega_0 (h_1^2 + h_2^2)}, \quad \beta = \frac{\omega_0 h_1^2}{h_1^2 h_2^2 + \omega_0 (h_1^2 + h_2^2)}, \quad \varkappa = \frac{h_1^2 h_2^2}{h_1^2 h_2^2 + \omega_0 (h_1^2 + h_2^2)}. \qquad (35)$$

Comme α, β, $\varkappa > 0$ et $\alpha + \beta + \varkappa = 1$, le compte suivant les formules (32) et (33) s'avère stable. Le calcul élémentaire des opérations arithmétiques pour l'algorithme (31)-(35) donne $Q_+ = 10\,(N_1 - 1) \times \times (N_2 - 1)$ opérations d'addition et de soustraction et $Q_* = Q_+$ opérations de multiplication, en tout $Q = 20\,(N_1 - 1)\,(N_2 - 1)$ opérations.

Compte tenu de l'estimation trouvée auparavant pour le nombre d'itérations, on constate que pour le calcul de la solution du problème de différences (27) suivant l'algorithme (31)-(35) à la précision ε il faut dépenser, au cas de $N_1 = N_2 = N$, $l_1 = l_2 = l$,

$$Q(\varepsilon) \approx 5{,}6 N^2 \sqrt{N} \ln(2/\varepsilon)$$

opérations arithmétiques.

Voyons maintenant le premier algorithme qui, dans le cas considéré, a la forme

$$\begin{gathered} \varphi_k = (E + \omega_0 R_1)(E + \omega_0 R_2)\, y_k - \tau_{k+1}(A y_k - f), \\ (E + \omega_0 R_1)\, v = \varphi_k, \quad (E + \omega_0 R_2)\, y_{k+1} = v. \end{gathered} \qquad (36)$$

Avec cet algorithme, si l'on passe à l'écriture discrète de ce dernier, il sera plus commode de recourir aux fonctions de mailles associées à $\bar{\omega}$ et s'annulant sur γ. Ces fonctions coïncident avec y_k, v et y_{k+1} sur ω et, comme d'habitude, sont désignées par $\mathring{y}_k$, $\mathring{v}$ et $\mathring{y}_{k+1}$. Pour obtenir l'approximation à la solution du problème (27), on doit la définir ainsi: $y_k(x) = \mathring{y}_k(x)$ pour $x \in \omega$ et $y_k(x) = g(x)$, $x \in \gamma$.

Afin de réaliser dans (36) le passage à l'écriture aux différences (par points), il est nécessaire de déterminer l'opérateur de différences $\bar{\mathscr{R}}$ correspondant au produit des opérateurs $R_1 R_2$. Notons qu'en vertu de la définition les opérateurs R_1 et R_2 s'écrivent de la sorte:

$$R_1 y = \begin{cases} \dfrac{1}{h_1} y_{\bar{x}_1} + \dfrac{1}{h_2} y_{\bar{x}_2}, & 2 \leqslant i \leqslant N_1 - 1, \quad 2 \leqslant j \leqslant N_2 - 1, \\ \dfrac{1}{h_1^2} y + \dfrac{1}{h_2} y_{\bar{x}_2}, & i = 1, \quad 2 \leqslant j \leqslant N_2 - 1, \\ \dfrac{1}{h_1} y_{\bar{x}_1} + \dfrac{1}{h_2^2} y, & 2 \leqslant i \leqslant N_1 - 1, \quad j = 1, \\ \left(\dfrac{1}{h_1^2} + \dfrac{1}{h_2^2}\right) y, & i = j = 1, \end{cases}$$

$$R_2 y = \begin{cases} -\dfrac{1}{h_1} y_{x_1} - \dfrac{1}{h_2} y_{x_2}, & 1 \leqslant i \leqslant N_1 - 2, \quad 1 \leqslant j \leqslant N_2 - 2, \\ \dfrac{1}{h_1^2} y - \dfrac{1}{h_2} y_{x_2}, & i = N_1 - 1, \quad 1 \leqslant j \leqslant N_2 - 2, \\ -\dfrac{1}{h_1} y_{x_1} + \dfrac{1}{h_2^2} y, & 1 \leqslant i \leqslant N_1 - 2, \quad j = N_2 - 1, \\ \left(\dfrac{1}{h_1^2} + \dfrac{1}{h_2^2}\right) y, & i = N_1 - 1, \quad j = N_2 - 1. \end{cases}$$

Les calculs montrent que si l'opérateur $\overline{\mathscr{R}}$ est défini de la façon suivante :

$$\overline{\mathscr{R}}y=\sum_{\alpha=1}^{2}\frac{1}{h_\alpha^2}y_{\bar{x}_\alpha x_\alpha}+\frac{1}{h_1h_2}(y_{x_2\bar{x}_1}+y_{x_1\bar{x}_2})+qy,\qquad x\in\omega,$$

où

$$q(i,j)=\begin{cases}0, & 2\leqslant i\leqslant N_1-1,\quad 2\leqslant j\leqslant N_2-1,\\ \dfrac{1}{h_1^4}, & i=1,\quad 2\leqslant j\leqslant N_2-1,\\ \dfrac{1}{h_2^4}, & 2\leqslant i\leqslant N_1-1,\quad j=1,\\ \dfrac{1}{h_1^4}+\dfrac{1}{h_2^4}, & i=1,\quad j=1,\end{cases}$$

alors $R_1R_2y=-\overline{\mathscr{R}}\overset{\circ}{y}$, où $y\in H$, $\overset{\circ}{y}\in\overset{\circ}{H}$ et $y(x)=\overset{\circ}{y}(x)$ pour $x\in\omega$.

Utilisons la définition des opérateurs A, R_1 et R_2, ainsi que l'expression obtenue pour R_1R_2, et écrivons l'algorithme (36) de la sorte

$$\begin{aligned}\varphi_k(i,j)=[d_{k+1}-q(i,j)\,\omega_0^2]\,\overset{\circ}{y}_k(i,j)+a_{k+1}[\overset{\circ}{y}_k(i+1,j)+\\+\overset{\circ}{y}_k(i-1,j)]+b_{k+1}[\overset{\circ}{y}_k(i,j+1)+\overset{\circ}{y}_k(i,j-1)]+\\+c\,[\overset{\circ}{y}_k(i-1,j+1)+\overset{\circ}{y}_k(i+1,j-1)]+\tau_{k+1}f(i,j),\end{aligned}\qquad(37)$$

avec les notations

$$a_{k+1}=\frac{\tau_{k+1}}{h_1^2}-\frac{\omega_0}{h_1^2}\left(1+\frac{\omega_0}{h_1^2}+\frac{\omega_0}{h_2^2}\right),$$

$$b_{k+1}=\frac{\tau_{k+1}}{h_2^2}-\frac{\omega_0}{h_2^2}\left(1+\frac{\omega_0}{h_1^2}+\frac{\omega_0}{h_2^2}\right),$$

$$c=\frac{\omega_0^2}{h_1^2h_2^2},\qquad d_{k+1}=1-2\,(a_{k+1}+b_{k+1}+c).$$

Ensuite,

$$\begin{aligned}&v(i,j)=\alpha v(i-1,j)+\beta v(i,j-1)+\varkappa\varphi_k(i,j),\\&i=1,2,\ldots,N_1-1,\quad j=1,2,\ldots,N_2-1,\\&v(0,j)=0,\quad 1\leqslant j\leqslant N_2-1,\quad v(i,0)=0,\quad 1\leqslant i\leqslant N_1-1,\end{aligned}\qquad(38)$$

$$\begin{aligned}&\overset{\circ}{y}_{k+1}(i,j)=\alpha\overset{\circ}{y}_{k+1}(i+1,j)+\beta\overset{\circ}{y}_{k+1}(i,j+1)+\varkappa v(i,j),\\&i=N_1-1,N_1-2,\ldots,1,\quad j=N_2-1,N_2-2,\ldots,1,\\&\overset{\circ}{y}_{k+1}|_\gamma=0,\end{aligned}\qquad(39)$$

où α, β et $\varkappa$ sont définis dans (35). Le calcul du nombre d'opérations arithmétiques donne $Q_+=11N_1N_2-10(N_1+N_2)+10$ additions et soustractions et $Q_*=Q_+$ multiplications, en tout $Q=$ $=22N_1N_2-20(N_1+N_2)+20$ opérations. C'est environ 1,1 fois

plus grand que pour l'algorithme (31)-(34). L'avantage de l'algorithme (37)-(39) consiste ainsi dans le fait que dans le cas considéré il ne faut pas mémoriser l'information intermédiaire sur $\varphi_k(i, j)$, $v(i, j)$, et $\overset{\circ}{y}_{k+1}(i, j)$, déterminé de nouveau, occupent, au fur et à mesure, la place prise par $\overset{\circ}{y}_k(i, j)$.

R e m a r q u e 2. Au cas de p dimensions, l'opérateur $\overline{\mathcal{R}}$ prend la forme

$$\overline{\mathcal{R}}y = \sum_{\alpha=1}^{p} \frac{1}{h_\alpha^2} y_{\bar{x}_\alpha x_\alpha} + \sum_{\alpha=1}^{p} \sum_{\beta\neq\alpha}^{1\div p} \frac{1}{h_\alpha^2 h_\beta^2} y_{x_\beta \bar{x}_\beta \bar{x}_\alpha x_\alpha} + qy,$$

où

$$q(i_1, i_2, \ldots, i_p) = \sum_{\alpha=1}^{p} \frac{\delta_{i_\alpha, 1}}{h_\alpha^4}, \quad \delta_{i,j} = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

R e m a r q u e 3. A la méthode triangulaire alternée par blocs correspond la définition suivante des opérateurs de différences $\mathcal{R}_1$ et $\mathcal{R}_2$:

$$\mathcal{R}_1 y = -\frac{1}{2} y_{\bar{x}_1 x_1} - \frac{1}{h_2} y_{\bar{x}_2}, \quad \mathcal{R}_2 y = -\frac{1}{2} y_{\bar{x}_1 x_1} + \frac{1}{h_2} y_{x_2}.$$

Dans ce cas pour l'inversion de l'opérateur B il faut recourir à la méthode du balayage triponctuel. Les calculs deviennent alors plus laborieux pour chaque itération et ce travail n'est pas compensé par une faible diminution du nombre d'itérations (environ de 1,2 fois).

§ 2. Problèmes aux limites discrets pour les équations elliptiques dans un rectangle

1. Problème de Dirichlet pour équation à coefficients variables. Voyons maintenant comment s'applique la méthode triangulaire alternée à la recherche de la solution du problème discret de Dirichlet pour l'équation elliptique sans dérivées mixtes

$$\begin{aligned} \Lambda y = \sum_{\alpha=1}^{2} (a_\alpha(x) y_{\bar{x}_\alpha})_{x_\alpha} = -\varphi(x), \quad x \in \omega, \\ y(x) = g(x), \quad x \in \gamma, \end{aligned} \tag{1}$$

dans un rectangle, où $\bar{\omega} = \omega \cup \gamma$ est un maillage régulier rectangulaire de pas h_1 et h_2: $\bar{\omega} = \{x_{ij} = (ih_1, jh_2), 0 \leqslant i \leqslant N_1, 0 \leqslant j \leqslant N_2, h_\alpha N_\alpha = l_\alpha, \alpha = 1, 2\}$. Posons que les coefficients $a_\alpha(x)$ satisfont aux conditions

$$0 < c_1 \leqslant a_\alpha(x) \leqslant c_2, \quad \alpha = 1, 2. \tag{2}$$

Exigeons de même pour un j fixé, $1 \leqslant j \leqslant N_2 - 1$, que le nombre de nœuds du maillage ω, dans lesquels $(a_1)_{x_1} = O(h_1^{-1})$, soit fini et ne dépende pas de h_1. Cela signifie que le coefficient correspondant dans l'équation différentielle, pour chaque x_2 fixé, possède un nombre fini de points de discontinuité suivant la direction de x_1. Une exigence analogue doit concerner $(a_2)_{x_2}$.

Le problème de différences (1) se réduit à l'équation opératorielle

$$Au = f \tag{3}$$

de façon banale. H est ici l'espace des fonctions de mailles données sur ω avec produit scalaire

$$(u, v) = \sum_{x \in \omega} u(x) v(x) h_1 h_2,$$

$$Ay = -\Lambda \mathring{y}, \quad y \in H, \quad \mathring{y} \in \mathring{H} \text{ et } y(x) = \mathring{y}(x) \text{ pour } x \in \omega;$$

$$f(x) = \varphi(x) + \frac{1}{h_1^2} \varphi_1(x) + \frac{1}{h_1^2} \varphi_2(x),$$

où

$$\varphi_1(x) = \begin{cases} a_1(h_1, x_2)\, g(0, x_2), & x_1 = h_1, \\ 0, & 2h_1 \leqslant x_1 \leqslant l_1 - 2h_1, \\ a_1(l_1, x_2)\, g(l_1, x_2), & x_1 = l_1 - h_1, \end{cases}$$

$$\varphi_2(x) = \begin{cases} a_2(x_1, h_2)\, g(x_1, 0), & x_2 = h_2, \\ 0, & 2h_2 \leqslant x_2 \leqslant l_2 - 2h_2, \\ a_2(x_1, l_2)\, g(x_1, l_2), & x_2 = l_2 - h_2. \end{cases}$$

En utilisant les formules discrètes de Green, on trouve que l'opérateur A est autoadjoint dans H et qu'on a l'égalité

$$(Ay, y) = -(\Lambda \mathring{y}, \mathring{y}) = \sum_{\alpha=1}^{2} (a_\alpha \mathring{y}_{\bar{x}_\alpha}^2, 1)_\alpha, \tag{4}$$

où

$$(u, v)_\alpha = \sum_{x_\alpha = h_\alpha}^{l_\alpha} \sum_{x_\beta = h_\beta}^{l_\beta - h_\beta} u(x) v(x) h_\alpha h_\beta, \quad \beta = 3 - \alpha, \quad \alpha = 1, 2.$$

Pour la résolution approchée de l'équation, considérons la méthode triangulaire alternée construite avec l'utilisation du régularisateur $R \neq A$:

$$\begin{gathered} B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \ldots, y_0 \in H, \\ B = (E + \omega R_1)(E + \omega R_2), \quad R_1 = R_2^*, \quad R = R_1 + R_2. \end{gathered} \tag{5}$$

Le régularisateur R sera choisi de la façon suivante :

$$Ry = -\mathcal{R}\overset{\circ}{y}, \quad \mathcal{R}y = y_{\bar{x}_1x_1} + y_{\bar{x}_2x_2}, \quad \overset{\circ}{y} \in \overset{\circ}{H}, \tag{6}$$

tandis que les opérateurs R_1 et R_2 seront définis suivant les formules

$$R_\alpha y = -\mathcal{R}_\alpha \overset{\circ}{y}, \quad \mathcal{R}_1 y = -\sum_{\alpha=1}^{2} \frac{1}{h_\alpha} y_{\bar{x}_\alpha}, \quad \mathcal{R}_2 y = \sum_{\alpha=1}^{2} \frac{1}{h_\alpha} y_{x_\alpha}. \tag{7}$$

Au point 4 du § 1 on a montré que pour les opérateurs R_1 et R_2 définis ici on a les inégalités

$$\delta E \leqslant R, \quad R_1 R_2 \leqslant (\Delta/4) R,$$

$$\delta = \sum_{\alpha=1}^{2} \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2l_\alpha}, \quad \Delta = \sum_{\alpha=1}^{2} \frac{4}{h_\alpha^2}.$$

Ensuite, profitons des formules de différences de Green, il vient

$$(Ry, y) = -(\mathcal{R}\overset{\circ}{y}, \overset{\circ}{y}) = \sum_{\alpha=1}^{2} (\overset{\circ}{y}{}^2_{\bar{x}_\alpha}, 1)_\alpha. \tag{8}$$

Par conséquent, de (2), (4) et (8) découlent les inégalités $c_1 R \leqslant A \leqslant$ $\leqslant c_2 R$, $c_1 > 0$. Vu que l'opérateur R est autoadjoint et défini positif dans H, à la méthode considérée s'applique le théorème 2 avec $\mathcal{D} = E$, qui indique comment il faut choisir les paramètres d'itération ω et $\{\tau_k\}$. Ce théorème fournit également l'estimation de l'erreur

$$\| z_n \|_D \leqslant q_n \| z_0 \|_D, \quad D = A, \ B \text{ ou } AB^{-1}A,$$

où

$$q_n = \frac{2\rho_1^n}{1+\rho_1^{2n}}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{c_1}{c_2} \frac{2\sqrt{\eta}}{1+\sqrt{\eta}}, \quad \eta = \frac{\delta}{\Delta}.$$

Pour un η petit, on obtient l'estimation du nombre d'itérations :

$$n \geqslant n_0(\varepsilon), \quad n_0(\varepsilon) = \sqrt{\frac{c_2}{c_1}} \frac{\ln(2/\varepsilon)}{2\sqrt{2}\sqrt[4]{\eta}}, \quad \eta = \frac{\pi^2}{4N^2}.$$

Il s'ensuit que le nombre d'itérations est proportionnel à $\sqrt{c_2/c_1}$ et qu'il est rationnel d'utiliser la méthode (5), (7) quand ce rapport n'est pas trop grand.

2. Méthode triangulaire alternée modifiée *). Poursuivons l'étude de la méthode triangulaire alternée appliquée au problème de différences (1) au cas où les coefficients $a_\alpha(x)$ varient fortement c'est-à-dire que le rapport c_2/c_1 est grand.

---

*) Voir A. B. Koutchérov et E. S. Nikolaïev (ЖВМ et МФ, 16, nº 5, 1976; 17, nº 3, 1977).

Voyons à présent pour l'équation (3) la variante modifiée de la méthode triangulaire alternée

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}}+Ay_k=f,\quad k=0,1,\ldots,$$
$$B=(\mathscr{D}+\omega R_1)\,\mathscr{D}^{-1}(\mathscr{D}+\omega R_2),\quad R_1=R_2^*,\quad R_1+R_2=A,\tag{9}$$

où l'on pose $\mathscr{D}y=d(x)y$, $x\in\omega$. $d(x)$ est ici une fonction de maille positive associée à ω et qui sera définie utlérieurement. Dans ce cas $\mathscr{D}$ est un opérateur autoadjoint et défini positif dans H. La fonction de maille $d(x)$ joue dans (9) le rôle de paramètre d'itération supplémentaire et permet ainsi de tenir compte des particularités de l'opérateur A en chaque nœud x du maillage ω.

Définissons maintenant les opérateurs R_α de la façon suivante: $R_\alpha y=-\mathscr{R}_\alpha\overset{\circ}{y}$, $y\in H$ et $\overset{\circ}{y}\in\overset{\circ}{H}$, où

$$\mathscr{R}_1y=-\sum_{\alpha=1}^{2}\left(\frac{a_\alpha}{h_\alpha}y_{\bar{x}_\alpha}+\frac{a_{\alpha x_\alpha}}{2h_\alpha}y\right),$$
$$\mathscr{R}_2y=\sum_{\alpha=1}^{2}\left(\frac{a_\alpha^{+1}}{h_\alpha}y_{x_\alpha}+\frac{a_{\alpha x_\alpha}}{2h_\alpha}y\right),\quad x\in\omega,\tag{10}$$

et $a_1^{\pm1}(x)=a_1(x_1\pm h_1,\ x_2)$, $a_2^{\pm1}(x)=a_2(x_1,\ x_2\pm h_2)$.

Montrons que les opérateurs R_1 et R_2 sont des opérateurs autoadjoints dans H. Il suffit pour cela de montrer que l'égalité $(\mathscr{R}_1\overset{\circ}{y},\overset{\circ}{v})=(\overset{\circ}{y},\mathscr{R}_2\overset{\circ}{v})$, $\overset{\circ}{y}\in\overset{\circ}{H}$, $\overset{\circ}{v}\in\overset{\circ}{H}$ est vérifiée. Il résulte des formules de différences de Green pour des fonctions, s'annulant sur γ, et de la formule de la dérivation au sens des différences finies du produit des fonctions de mailles $(yv)_{x_\alpha}=y^{+1}v_{x_\alpha}+y_{x_\alpha}v$ que

$$(\mathscr{R}_1\overset{\circ}{y},\overset{\circ}{v})=-\sum_{\alpha=1}^{2}\frac{1}{h_\alpha}(a_\alpha\overset{\circ}{y}_{\bar{x}_\alpha},\overset{\circ}{v})-\sum_{\alpha=1}^{2}\frac{1}{2h_\alpha}(a_{\alpha x_\alpha}\overset{\circ}{y},\overset{\circ}{v})=$$
$$=\sum_{\alpha=1}^{2}\left[\frac{1}{h_\alpha}(\overset{\circ}{y},(a_\alpha\overset{\circ}{v})_{x_\alpha})-\frac{1}{2h_\alpha}(a_{\alpha x_\alpha}\overset{\circ}{y},\overset{\circ}{v})\right]=$$
$$=\sum_{\alpha=1}^{2}\left[\frac{1}{h_\alpha}(\overset{\circ}{y},a_\alpha^{+1}\overset{\circ}{v}_{x_\alpha})+\frac{1}{2h_\alpha}(\overset{\circ}{y},a_{\alpha x_\alpha}\overset{\circ}{v})\right]=(\overset{\circ}{y},\mathscr{R}_2\overset{\circ}{v}).$$

La proposition est démontrée.

Comme $R_1+R_2=A$, suivant le théorème 1 l'information à priori pour la méthode triangulaire alternée (9) est de la forme des

constantes δ et Δ des inégalités

$$\delta\mathscr{D}\leqslant A, \quad R_1\mathscr{D}^{-1}R_2\leqslant\frac{\Delta}{4}A, \quad \delta>0. \tag{11}$$

Vu que le rapport $\eta=\delta/\Delta$ définit le nombre d'itérations, la fonction de maille $d(x)$ doit être choisie sur la base de la condition du maximum de ce rapport.

Passons à présent au choix de la fonction $d(x)$ et aux estimations de δ et Δ. Démontrons d'abord une inégalité.

Lemme 2. *Soient $p_\alpha(x)$, $q_\alpha(x)$, $u_\alpha(x)$ et $v_\alpha(x)$, $\alpha=1, 2$, les fonctions de mailles données sur ω. Alors pour tout $x\in\omega$ on a l'inégalité*

$$\Big[\sum_{\alpha=1}^{2}(p_\alpha u_\alpha+q_\alpha v_\alpha)\Big]^2\leqslant$$
$$\leqslant(1+\varepsilon)(|p_1|+\varkappa_1|q_1|)\Big(|p_1|u_1^2+\frac{|q_1|}{\varkappa_1}v_1^2\Big)+$$
$$+\frac{1+\varepsilon}{\varepsilon}(|p_2|+\varkappa_2|q_2|)\Big(|p_2|u_2^2+\frac{|q_2|}{\varkappa_2}v_2^2\Big), \tag{12}$$

où $\varepsilon(x)$, $\varkappa_1(x)$ et $\varkappa_2(x)$ sont des fonctions de mailles positives quelconques associées à ω.

En effet, en utilisant ε et l'inégalité $2ab\leqslant\varepsilon a^2+b^2/\varepsilon$, $\varepsilon>0$, il vient

$$\Big[\sum_{\alpha=1}^{2}(p_\alpha u_\alpha+q_\alpha v_\alpha)\Big]^2=$$
$$=(p_1u_1+q_1v_1)^2+2(p_1u_1+q_1v_1)(p_2u_2+q_2v_2)+(p_2u_2+q_2v_2)^2\leqslant$$
$$\leqslant(1+\varepsilon)(p_1u_1+q_1v_1)^2+\frac{1+\varepsilon}{\varepsilon}(p_2u_2+q_2v_2)^2. \tag{13}$$

En recourant de nouveau à l'inégalité mentionnée, on obtient

$$(p_\alpha u_\alpha+q_\alpha v_\alpha)^2=p_\alpha^2u_\alpha^2+2p_\alpha q_\alpha u_\alpha v_\alpha+q_\alpha^2v_\alpha^2\leqslant$$
$$\leqslant p_\alpha^2u_\alpha^2+|p_\alpha||q_\alpha|\Big(\varkappa_\alpha u_\alpha^2+\frac{1}{\varkappa_\alpha}v_\alpha^2\Big)+q_\alpha^2v_\alpha^2=$$
$$=(|p_\alpha|+\varkappa_\alpha|q_\alpha|)\Big(|p_\alpha|u_\alpha^2+\frac{|q_\alpha|}{\varkappa_\alpha}v_\alpha^2\Big), \quad \varkappa_\alpha>0, \quad \alpha=1, 2.$$

En portant l'inégalité obtenue dans (13), on aboutira à (12). Le lemme est démontré.

Profitons de l'inégalité (12), ainsi que de la définition des opérateurs R_1 et R_2, et l'on trouve que

$$(R_1\mathscr{D}^{-1}R_2y,\ y)=(\mathscr{D}^{-1}R_2\overset{\circ}{y},\ R_2\overset{\circ}{y})=$$

$$=\Big(\frac{1}{d}\sum_{\alpha=1}^{2}\Big(\frac{a_\alpha^{+1}}{h_\alpha}\overset{\circ}{y}_{x_\alpha}+\frac{a_{\alpha x_\alpha}}{2h_\alpha}\overset{\circ}{y}\Big)^2,\ 1\Big)\leqslant$$

$$\leqslant\Big(\frac{(1+\varepsilon)}{dh_1^2}(a_1^{+1}+0{,}5h_1\varkappa_1\,|a_{1x_1}|)\Big(a_1^{+1}\overset{\circ}{y}{}_{\bar{x}_1}^2+\frac{0{,}5h_1\,|a_{1x_1}|}{\varkappa_1h_1^2}\overset{\circ}{y}{}^2\Big),\ 1\Big)+$$

$$+\Big(\frac{(1+\varepsilon)}{d\varepsilon h_2^2}(a_2^{+1}+0{,}5h_2\varkappa_2\,|a_{2x_2}|)\Big(a_2^{+1}\overset{\circ}{y}{}_{\bar{x}_2}^2+\frac{0{,}5h_2\,|a_{2x_2}|}{\varkappa_2h_2^2}\overset{\circ}{y}{}^2\Big),\ 1\Big).$$

Notons que dans (12) à la place de p_α, q_α, u_α et v_α se trouvent $p_\alpha=\frac{a_\alpha^{+1}}{h_\alpha}$, $q_\alpha=0{,}5a_{\alpha x_\alpha}$, $u_\alpha=\overset{\circ}{y}_{x_\alpha}$, $v_\alpha=\frac{1}{h_\alpha}\overset{\circ}{y}$, $\alpha=1,\ 2$. Exigeons que dans l'inégalité obtenue $\varkappa_1$ ne soit une fonction que de x_2, tandis que $\varkappa_2$ le soit uniquement de x_1, autrement dit posons

$$\varkappa_\alpha=\varkappa_\alpha(x_\beta),\quad \beta=3-\alpha,\quad \alpha=1,\ 2. \tag{14}$$

Posons

$$\varepsilon=\varepsilon(x)=\frac{a_2^{+1}+0{,}5h_2\varkappa_2\,|a_{2x_2}|}{a_1^{+1}+0{,}5h_1\varkappa_1\,|a_{1x_1}|}\cdot\frac{h_1^2\theta_2(x_1)}{h_2^2\theta_1(x_2)} \tag{15}$$

et définissons $d(x)$ de la façon suivante:

$$d(x)=\sum_{\alpha=1}^{2}(a_\alpha^{+1}+0{,}5h_\alpha\varkappa_\alpha\,|a_{\alpha x_\alpha}|)\frac{\theta_\alpha}{h_\alpha^2}, \tag{16}$$

où $\theta_\alpha=\theta_\alpha(x_\beta)$, $\beta=3-\alpha$, $\alpha=1,\ 2$ sont des fonctions de mailles positives associées à ω et qui doivent être définies.

En portant (15) et (16) dans l'inégalité obtenue auparavant, il vient

$$(R_1\mathscr{D}^{-1}R_2y,\ y)\leqslant\sum_{\alpha=1}^{2}\Big(\frac{a_\alpha^{+1}}{\theta_\alpha}\overset{\circ}{y}{}_{\bar{x}_\alpha}^2,\ 1\Big)+\sum_{\alpha=1}^{2}\Big(\frac{|a_{\alpha x_\alpha}|}{2h_\alpha\theta_\alpha\varkappa_\alpha}\overset{\circ}{y}{}^2,\ 1\Big).$$

Vu que θ_α ne dépend pas de x_α, en utilisant le produit scalaire $(,)_\alpha$ introduit avant, on aboutit à ce que

$$\Big(\frac{a_\alpha^{+1}}{\theta_\alpha}\overset{\circ}{y}{}_{\bar{x}_\alpha}^2,\ 1\Big)\leqslant\Big(\frac{a_\alpha}{\theta_\alpha}\overset{\circ}{y}{}_{\bar{x}_\alpha}^2,\ 1\Big)_\alpha,\quad \alpha=1,\ 2.$$

Par conséquent,

$$(R_1\mathscr{D}^{-1}R_2y,\ y)\leqslant\sum_{\alpha=1}^{2}\Big(\frac{a_\alpha}{\theta_\alpha}\overset{\circ}{y}{}_{\bar{x}_\alpha}^2,\ 1\Big)_\alpha+\sum_{\alpha=1}^{2}\Big(\frac{|a_{\alpha x_\alpha}|}{2h_\alpha\theta_\alpha\varkappa_\alpha}\overset{\circ}{y}{}^2,1\Big). \tag{17}$$

Choisissons maintenant θ_α et $\varkappa_\alpha$. Posons

$$\omega_1 = \{x_1 = ih_1, \quad 1 \leqslant i \leqslant N_1 - 1, \quad h_1 N_1 = l_1\},$$
$$\omega_1^+ = \{x_1 = ih_1, \quad 1 \leqslant i \leqslant N_1, \quad h_1 N_1 = l_1\}$$

et déterminons

$$(u, v)_{\omega_1} = \sum_{x_1 \in \omega_1} u(x) v(x) h_1,$$
$$(u, v)_{\omega_1^+} = \sum_{x_1 \in \omega_1^+} u(x) v(x) h_1.$$

De façon analogue sont introduits ω_2 et ω_2^+ ainsi que $(u, v)_{\omega_2}$ et $(u, v)_{\omega_2^+}$. On voit alors sans peine qu'on a des relations

$$\begin{gathered}(u, v) = ((u, v)_{\omega_1}, 1)_{\omega_2} = ((u, v)_{\omega_2}, 1)_{\omega_1},\\ (u, v)_\alpha = ((u, v)_{\omega_\alpha^+}, 1)_{\omega_\beta}, \quad \beta = 3 - \alpha; \quad \alpha = 1, 2.\end{gathered} \tag{18}$$

Soit à présent $b_\alpha(x_\beta) = \max\limits_{x_\alpha \in \omega_\alpha} v^\alpha(x)$, $\alpha = 1, 2$, $x_\beta \in \omega_\beta$, où $v^\alpha(x)$ pour un x_β fixé est la solution du problème aux limites triponctuel suivant :

$$\begin{gathered}(a_\alpha v^\alpha_{\bar{x}_\alpha})_{x_\alpha} = -\frac{a_\alpha^{+1}}{h_\alpha^2} \quad h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha,\\ v^\alpha(x) = 0, \quad x_\alpha = 0, l_\alpha, \quad x_\beta \in \omega_\beta.\end{gathered} \tag{19}$$

Alors, en vertu du lemme 13 (point 4, § 2, ch. V) on obtient

$$\left(\frac{a_\alpha^{+1}}{h_\alpha^2} \mathring{y}^2, 1\right)_{\omega_\alpha} \leqslant b_\alpha(x_\beta) (a_\alpha \mathring{y}^2_{\bar{x}_\alpha}, 1)_{\omega_\alpha^+}, \quad \alpha = 1, 2.$$

Multiplions cette inégalité par $\theta_\alpha(x_\beta)$ et sommons scalairement en ω_β. Alors, en vertu de (18), on a

$$\left(\frac{\theta_\alpha a_\alpha^{+1}}{h_\alpha^2} \mathring{y}^2, 1\right) \leqslant (b_\alpha a_\alpha \theta_\alpha \mathring{y}^2_{\bar{x}_\alpha}, 1)_\alpha, \quad \alpha = 1, 2. \tag{20}$$

Soit $c_\alpha(x_\beta) = \max\limits_{x_\alpha \in \omega_\alpha} w^\alpha(x)$, $\alpha = 1, 2$, $x_\beta \in \omega_\beta$, où $w^\alpha(x)$ pour un x_β fixé est la solution du problème aux limites triponctuel suivant :

$$\begin{gathered}(a_\alpha w^\alpha_{\bar{x}_\alpha})_{x_\alpha} = -\frac{|a_{\alpha x_\alpha}|}{2h_\alpha}, \quad h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha,\\ w^\alpha(x) = 0, \quad x_\alpha = 0, l_\alpha, \quad x_\beta \in \omega_\beta.\end{gathered} \tag{21}$$

De façon analogue à ce qu'on a obtenu les inégalités (20), on aboutit, en vertu de (14), aux inégalités suivantes:

$$\left(\frac{\varkappa_\alpha\theta_\alpha\,|a_{\alpha x_\alpha}|}{2h_\alpha}\,\overset{\circ}{y}{}^2,\ 1\right)\leqslant(\varkappa_\alpha\theta_\alpha c_\alpha a_\alpha \overset{\circ}{y}{}^2_{\bar{x}_\alpha},\ 1)_\alpha,\qquad \alpha=1,\ 2,\qquad (22)$$

$$\left(\frac{|a_{\alpha x_\alpha}|}{2h_\alpha\theta_\alpha\varkappa_\alpha}\,\overset{\circ}{y}{}^2,\ 1\right)\leqslant\left(\frac{c_\alpha}{\varkappa_\alpha\theta_\alpha}\,a_\alpha \overset{\circ}{y}{}^2_{\bar{x}_\alpha},\ 1\right)_\alpha,\qquad \alpha=1,\ 2.\qquad (23)$$

Additionnons maintenant les inégalités (20) et (22) et sommons-les en α. Alors, en vertu de (16), il vient

$$(d\overset{\circ}{y}{}^2,\ 1)=(\mathscr{D}y,\ y)\leqslant((\varkappa_\alpha c_\alpha+b_\alpha)\,\theta_\alpha a_\alpha \overset{\circ}{y}{}^2_{\bar{x}_\alpha},\ 1)_\alpha.$$

En choisissant θ_α au moyen de la formule

$$\theta_\alpha(x_\beta)=\frac{1}{b_\alpha(x_\beta)+c_\alpha(x_\beta)\,\varkappa_\alpha(x_\beta)},\qquad \beta=3-\alpha,\qquad \alpha=1,\ 2,\qquad (24)$$

et compte tenu de (4), on en tire que $(\mathscr{D}y,\ y)\leqslant(Ay,\ y)$. Par conséquent, on peut poser dans (11) $\delta=1$.

Apprécions maintenant Δ. A cette fin portons (23) dans (17) et tenons compte du choix de θ_α suivant la formule (24). On obtient finalement l'estimation suivante:

$$(R_1\mathscr{D}^{-1}R_2y,\ y)\leqslant\sum_{\alpha=1}^{2}((1+c_\alpha/\varkappa_\alpha)\,(b_\alpha+c_\alpha\varkappa_\alpha)\,a_\alpha \overset{\circ}{y}{}^2_{\bar{x}_\alpha},\ 1)_\alpha.$$

Choisissons maintenant $\varkappa_\alpha$ optimal sur la base de la condition du minimum de l'expression $(1+c_\alpha/\varkappa_\alpha)(b_\alpha+c_\alpha\varkappa_\alpha)$ en $\varkappa_\alpha$. On obtient $\varkappa_\alpha(x_\beta)=\sqrt{b_\alpha(x_\beta)}$, $\beta=3-\alpha$, $\alpha=1,\ 2$ avec

$$(R_1\mathscr{D}^{-1}R_2y,\ y)\leqslant\sum_{\alpha=1}^{2}((c_\alpha+\sqrt{b_\alpha})^2\,a_\alpha \overset{\circ}{y}{}^2_{\bar{x}_\alpha},\ 1).$$

En comparant cette estimation à (4), on obtient que dans les inégalités (11) on peut poser

$$\Delta=4\max_{\alpha=1,\ 2}\left(\max_{x_\beta\in\omega_\beta}(c_\alpha(x_\beta)+\sqrt{b_\alpha(x_\beta)})^2\right),\qquad \beta=3-\alpha.\qquad (25)$$

En portant dans (16) les expressions trouvées pour $\varkappa_\alpha$ et θ_α, on obtient pour la fonction $d(x)$ la représentation de la forme

$$d(x)=\sum_{\alpha=1}^{2}\left(\frac{a_\alpha^{+1}}{h_\alpha^2\sqrt{b_\alpha}}+\frac{|a_{\alpha x_\alpha}|}{2h_\alpha}\right)\frac{1}{c_\alpha+\sqrt{b_\alpha}},\qquad x\in\omega.\qquad (26)$$

Bref, on a obtenu $d(x)$ et les constantes δ et Δ. Il ne reste qu'à appliquer le théorème 1. Étant donné que $\delta=1$, $\omega_0=2/\sqrt{\Delta}$, et pour le

nombre d'itérations se vérifie l'estimation

$$n \geqslant n_0(\varepsilon), \quad n_0(\varepsilon) = \frac{\sqrt[4]{\Delta}\ln(2/\varepsilon)}{2\sqrt{2}}.$$

Ensuite, en vertu des conditions (2), on obtient à partir de (19) et (21) $b_\alpha = O(1/h_\alpha^2)$ et $c_\alpha = O(1/h_\alpha)$, si le nombre de points auxquels $a_{\alpha x_\alpha} = O(h_\alpha^{-1})$ est fini. Il en résulte que $n_0(\varepsilon) = O(\sqrt{N}\ln(2/\varepsilon))$.

Arrêtons-nous à présent sur la mise en œuvre de la variante construite de la méthode triangulaire alternée (9). D'abord pour un x_β fixé, $h_\beta \leqslant x_\beta \leqslant l_\beta - h_\beta$ on résout par la méthode du balayage les problèmes aux limites triponctuels (19) et (21) et on trouve les valeurs de $b_\alpha(x_\beta)$ et $c_\alpha(x_\beta)$, $\alpha = 1, 2$. Ces quatre fonctions de mailles unidimensionnelles sont mémorisées et utilisées au cours des itérations pour le calcul de $d(x)$ suivant la formule (26). La simplicité de la formule (26) permet de ne pas retenir la fonction de maille bidimensionnelle $d(x)$ et de la calculer de nouveau au fur et à mesure des besoins.

Ensuite, suivant la formule (25) on cherche Δ et l'on pose $\delta = 1$. Les valeurs des paramètres d'itération ω et τ_k seront déterminées pour le schéma (9) sur la base du théorème 1.

Pour la recherche de y_{k+1} sur la base de y_k fixé, on recourt au premier des algorithmes décrits au point 1, § 1 pour la méthode triangulaire alternée

$$(\mathscr{D} + \omega_0 R_1) v = \varphi_k, \quad (\mathscr{D} + \omega_0 R_2) y_{k+1} = \mathscr{D} v,$$
$$\varphi_k = (\mathscr{D} + \omega_0 R_1) \mathscr{D}^{-1} (\mathscr{D} + \omega_0 R_2) y_k - \tau_{k+1}(Ay_k - f). \quad (27)$$

Sans traîner sur les détails, donnons la forme discrète de l'algorithme (27):

$$v(i, j) = \alpha_1(i, j) v(i-1, j) + \beta_1(i, j) v(i, j-1) +$$
$$+ \varkappa(i, j)\varphi_k(i, j),$$
$$i = 1, 2, \ldots, N_1 - 1, \quad j = 1, 2, \ldots, N_2 - 1, \quad (28)$$
$$v(0, j) = 0, \quad 1 \leqslant j \leqslant N_2 - 1, \quad v(i, 0) = 0, \quad 1 \leqslant i \leqslant N_1 - 1.$$

$$\mathring{y}_{k+1}(i, j) = \alpha_2(i, j)\mathring{y}_{k+1}(i+1, j) + \beta_2(i, j)\mathring{y}_{k+1}(i, j+1) +$$
$$+ \varkappa(i, j) d(i, j) v(i, j), \quad i = N_1 - 1, \ldots, 1, \; j = N_2 - 1, \ldots, 1, \quad (29)$$

où

$$\alpha_1 = \frac{\omega_0 a_1 \varkappa}{h_1^2}, \quad \beta_1 = \frac{\omega_0 a_2 \varkappa}{h_2^2}, \quad \alpha_2 = \frac{\omega_0 a_1^{+1} \varkappa}{h_1^2}, \quad \beta_2 = \frac{\omega_0 a_2^{+1} \varkappa}{h_2^2},$$
$$\frac{1}{\varkappa} = d + \omega_0 \left[\frac{a_1^{+1} + a_1}{2h_1^2} + \frac{a_2^{+1} + a_2}{2h_2^2} \right].$$

Le second membre $\varphi_k(i, j)$ se calcule suivant les formules

$$\varphi_k(i, j) = [P(i-1, j) + Q(i, j-1) + S(i, j)]\,\overset{\circ}{y}_k(i, j) + \\ + R_1(i, j)\overset{\circ}{y}_k(i+1, j) + R_1(i-1, j)\,\overset{\circ}{y}_k(i-1, j) + \\ + R_2(i, j)\overset{\circ}{y}_k(i, j+1) + R_2(i, j-1)\,\overset{\circ}{y}_k(i, j-1) + \\ + G(i-1, j)\,\overset{\circ}{y}_k(i-1, j+1) + G(i, j-1)\,\overset{\circ}{y}_k(i+1, j-1) + \\ + \tau_{k+1} f(i, j),$$

$$1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant j \leqslant N_2 - 1, \tag{30}$$

où

$$G = \frac{\omega_0^2 a_1^{+1} a_2^{+1}}{h_1^2 h_2^2 d}, \quad R_\alpha = \left(\tau_{k+1} - \frac{\omega_0}{d\varkappa}\right)\frac{a_\alpha^{+1}}{h_\alpha^2}, \quad \alpha = 1, 2,$$

$$S = \frac{1}{\omega_0 \varkappa}\left[\frac{\omega_0}{d\varkappa} - 2\tau_{k+1}(1 - \varkappa d)\right], \quad P = \frac{\omega_0^2 (a_1^{+1})^2}{h_1^4 d}, \quad Q = \frac{\omega_0^2 (a_2^{+1})^2}{h_2^4 d},$$

avec $P(0, j) = 0$, $1 \leqslant j \leqslant N_2 - 1$, $Q(i, 0) = 0$, $1 \leqslant i \leqslant N_1 - 1$. Remarquons qu'en vertu de (25) et (26) les estimations

$$c_\alpha + \sqrt{b_\alpha} \leqslant \frac{\sqrt{\Delta}}{2} = \frac{1}{\omega_0}, \quad d \geqslant \omega_0 \sum_{\alpha=1}^{2} \frac{|a_{\alpha x_\alpha}|}{2h_\alpha}$$

se vérifient. De là on obtient

$$\frac{1}{\varkappa} = d + \omega_0 \sum_{\alpha=1}^{2} \frac{a_\alpha^{+1} + a_\alpha}{2h_\alpha^2} \geqslant \omega_0 \sum_{\alpha=1}^{2}\left(\frac{|a_{\alpha x_\alpha}|}{2h_\alpha} + \frac{a_\alpha^{+1} + a_\alpha}{2h_\alpha^2}\right) = \\ = \omega_0\left(\max\left(\frac{a_1^{+1}}{h_1^2}, \frac{a_1}{h_1^2}\right) + \max\left(\frac{a_2^{+1}}{h_2^2}, \frac{a_2}{h_2^2}\right)\right)$$

ou

$$\max(\alpha_1, \alpha_2) + \max(\beta_1, \beta_2) \leqslant 1.$$

Il s'ensuit que $\alpha_1 + \beta_1 \leqslant 1$ et $\alpha_2 + \beta_2 \leqslant 1$. Le calcul suivant les formules (28), (30) est donc stable.

3. Comparaison entre les différentes variantes de la méthode. Plus haut, en résolvant le problème (1), on a construit deux variantes de la méthode triangulaire alternée. La variante (5), (7) est bâtie sur la base du régularisateur R, tandis que la variante (9), (10) utilise l'opérateur $\mathscr{D}$ choisi de façon spéciale. Ces variantes possèdent la même caractéristique asymptotique établissant la dépendance du nombre d'itérations de celui des nœuds du maillage. Cependant l'estimation du nombre d'itérations de la première variante est fonction des caractéristiques extrémales des coefficients $a_\alpha(x)$, $\alpha = 1, 2$, de l'équation aux différences (1), tandis que pour la seconde variante elle est déterminée par leurs caractéristiques intégrales.

Comparons ces variantes de la méthode sur l'exemple modèle traditionnel. Soit donnée sur un maillage carré avec $N_1 = N_2 = N$, introduit dans un carré unitaire ($l_1 = l_2 = 1$), l'équation aux différences (1) dans laquelle

$$a_1(x) = 1 + c\,[(x_1 - 0{,}5)^2 + (x_2 - 0{,}5)^2],$$

$$a_2(x) = 1 + c\,[0{,}5 - (x_1 - 0{,}5)^2 - (x_2 - 0{,}5)^2], \quad x \in \overline{\omega}.$$

On a alors dans les inégalités (2) $c_1 = 1$, $c_2 = 1 + 0{,}5c$. En faisant varier le paramètre c, on obtiendra les coefficients $a_\alpha(x)$ aux propriétés extrémales différentes.

Tableau 10

c_2/c_1	$N = 32$		$N = 64$		$N = 128$	
	(5), (7)	(9), (10)	(5), (7)	(9), (10)	(5), (7)	(9), (10)
2	23	18	32	26	45	36
8	46	21	64	30	90	43
32	92	23	128	34	180	49
128	184	24	256	36	360	53
512	367	24	512	36	720	54

On a donné au tableau 10 le nombre d'itérations pour les variantes mentionnées en fonction du nombre de nœuds N suivant une direction et en fonction du rapport c_2/c_1 pour $\varepsilon = 10^{-4}$. On voit que pour des grandes valeurs de c_2/c_1 la méthode triangulaire alternée modifiée exige moins d'itérations, le nombre d'itérations dépendant faiblement de ce rapport.

4. Troisième problème aux limites. Etudions la méthode triangulaire alternée de résolution du troisième problème aux limites pour l'équation elliptique dans un rectangle $\overline{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha, \alpha = 1, 2\}$:

$$\sum_{\alpha=1}^{2} \frac{\partial}{\partial x_\alpha}\left(k_\alpha(x)\frac{\partial u}{\partial x_\alpha}\right) = -\varphi(x), \quad x \in G,$$

$$k_\alpha \frac{\partial u}{\partial x_\alpha} = \varkappa_{-\alpha}(x)\,u - g_{-\alpha}(x), \quad x_\alpha = 0, \tag{31}$$

$$-k_\alpha \frac{\partial u}{\partial x_\alpha} = \varkappa_{+\alpha}(x)\,u - g_{+\alpha}(x), \quad x_\alpha = l_\alpha.$$

Sur un maillage carré $\overline{\omega} = \{x_{ij} = (ih_1, jh_2) \in \overline{G},\ 0 \leqslant i \leqslant N_1,\ 0 \leqslant j \leqslant N_2,\ h_\alpha N_\alpha = l_\alpha,\ \alpha = 1, 2\}$ au problème (31) correspond le

problème de différences

$$\Lambda y = -f(x), \quad x \in \bar{\omega},$$
$$\Lambda = \Lambda_1 + \Lambda_2, \quad f(x) = \varphi(x) + \frac{2}{h_1}\varphi_1(x) + \frac{2}{h_2}\varphi_2(x), \tag{32}$$

où

$$\Lambda_\alpha y = \begin{cases} \frac{2}{h_\alpha}(a_\alpha^{+1} y_{x_\alpha} - \varkappa_{-\alpha} y), & x_\alpha = 0, \\ (a_\alpha y_{\bar{x}_\alpha})_{x_\alpha}, & h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ \frac{2}{h_\alpha}(-a_\alpha y_{\bar{x}_\alpha} - \varkappa_{+\alpha} y), & x_\alpha = l_\alpha, \end{cases}$$

$$\varphi_\alpha(x) = \begin{cases} g_{-\alpha}(x_\alpha), & x_\alpha = 0, \\ 0, & h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ g_{+\alpha}(x_\alpha), & x_\alpha = l_\alpha. \end{cases}$$

Admettons que les coefficients $a_\alpha(x)$ remplissent les conditions (2) et possèdent un nombre fini de points dans lesquels $a_{\alpha x_\alpha} = O(h_\alpha^{-1})$. Admettons également que $\varkappa_{-\alpha}(x_\beta)$ et $\varkappa_{+\alpha}(x_\beta)$ ne s'annulent pas simultanément pour chaque x_β fixé ($\varkappa_{-\alpha} \geqslant 0$, $\varkappa_{+\alpha} \geqslant 0$, $\varkappa_{-\alpha} + \varkappa_{+\alpha} > 0$).

Il est commode, au préalable, de réduire le problème de différences (32) au problème de Dirichlet dans un domaine étendu $\bar{\omega}^* = \{x_{ij} = (ih_1, jh_2), -1 \leqslant i \leqslant N_1 + 1, -1 \leqslant j \leqslant N_2 + 1\}$, pour lequel le maillage $\bar{\omega}$ est interne. Désignons par γ^* la frontière du maillage $\bar{\omega}^*$ et complétons la définition de la fonction de maille $y(x)$ d'un élément nul sur γ^*. Avec les notations

$$\bar{a}_\alpha(x) = \begin{cases} \rho(x_\beta) h_\alpha \varkappa_{-\alpha}(x_\beta), & x_\alpha = 0, \\ \rho(x_\beta) a_\alpha(x), & h_\alpha \leqslant x_\alpha \leqslant l_\alpha, \\ \rho(x_\beta) h_\alpha \varkappa_{+\alpha}(x_\beta), & x_\alpha = l_\alpha + h_\alpha, \quad 0 \leqslant x_\beta \leqslant l_\beta, \end{cases}$$

$$\bar{f}(x) = \rho(x_1)\rho(x_2) f(x), \quad x \in \bar{\omega},$$

$$\rho(x_\beta) = \begin{cases} 0{,}5, & x_\beta = 0, l_\beta, \\ 1, & h_\beta \leqslant x_\beta \leqslant l_\beta - h_\beta, \end{cases}$$

$$\beta = 3 - \alpha, \quad \alpha = 1, 2,$$

le problème (32) peut être écrit de la sorte

$$\bar{\Lambda} y = \sum_{\alpha=1}^{2} (\bar{a}_\alpha y_{\bar{x}_\alpha})_{x_\alpha} = -\bar{f}(x), \quad x \in \bar{\omega},$$
$$y(x) = 0, \quad x \in \gamma^*. \tag{33}$$

Rappelons que pour le problème de différences de la forme (33) on a bâti au point 2 la méthode triangulaire alternée modifiée (9)-(10).

Par conséquent, dans les formules du point 2 il ne faut que substituer $\bar{a}_\alpha(x)$ à $a_\alpha(x)$ pour obtenir la méthode de résolution du troisième problème aux limites pour l'équation elliptique dans un rectangle.

Dans le cas considéré, le problème aux limites triponctuel (19) s'écrit sous la forme

$$\begin{aligned} &(\bar{a}_\alpha v^\alpha_{\bar{x}_\alpha})_{x_\alpha} = -\frac{\bar{a}^{+1}_\alpha}{h^2_\alpha}, \quad 0 \leqslant x_\alpha \leqslant l_\alpha, \\ &v^\alpha(x) = 0, \quad x_\alpha = -h_\alpha,\ l_\alpha + h_\alpha. \end{aligned} \tag{34}$$

En utilisant les notations introduites plus haut pour $\bar{a}_\alpha$, on constate que (34) peut se mettre sous la forme

$$\begin{aligned} &\left(a_\alpha v^\alpha_{\bar{x}_\alpha}\right)_{x_\alpha} = -\frac{a^{+1}_\alpha}{h^2_\alpha}, \quad h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ &a^{+1}_\alpha v^\alpha_{x_\alpha} - \varkappa_{-\alpha} v^\alpha = -\frac{a^{+1}_\alpha}{h_\alpha}, \quad x_\alpha = 0, \\ &-a_\alpha v^\alpha_{\bar{x}_\alpha} - \varkappa_{+\alpha} v^\alpha = -\varkappa_{+\alpha}, \quad x_\alpha = l_\alpha. \end{aligned} \tag{35}$$

En vertu des hypothèses faites relativement à $a_\alpha(x)$, $\varkappa_{-\alpha}$ et $\varkappa_{+\alpha}$, le problème de différences a une solution. De plus,

$$b_\alpha(x_\beta) = \max_{0 \leqslant x_\alpha \leqslant l_\alpha} v^\alpha(x) = O\left(\frac{1}{h^2_\alpha}\right), \quad 0 \leqslant x_\beta \leqslant l_\beta.$$

De façon analogue, le problème (21), qui dans le cas concerné a pour expression

$$\begin{aligned} &\left(\bar{a}_\alpha w^\alpha_{\bar{x}_\alpha}\right)_{x_\alpha} = -\frac{|\bar{a}_{\alpha x_\alpha}|}{2h_\alpha}, \quad 0 \leqslant x_\alpha \leqslant l_\alpha, \\ &w^\alpha(x) = 0, \quad x_\alpha = -h_\alpha,\ l_\alpha + h_\alpha, \end{aligned}$$

en raison des notations, se réduit au troisième problème aux limites

$$\begin{aligned} &\left(a_\alpha w^\alpha_{\bar{x}_\alpha}\right)_{x_\alpha} = -\frac{|a_{\alpha x_\alpha}|}{2h_\alpha}, \quad h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ &a^{+1}_\alpha w^\alpha_{x_\alpha} - \varkappa_{-\alpha} w^\alpha = -\frac{|a^{+1}_\alpha - h_\alpha \varkappa_{-\alpha}|}{2h_\alpha}, \quad x_\alpha = 0, \\ &-a_\alpha w^\alpha_{\bar{x}_\alpha} - \varkappa_{+\alpha} w^\alpha = -\frac{|a_\alpha - h_\alpha \varkappa_{+\alpha}|}{2h_\alpha}, \quad x_\alpha = l_\alpha. \end{aligned} \tag{36}$$

De là on obtient

$$c_\alpha(x_\beta) = \max_{0 \leqslant x_\alpha \leqslant l_\alpha} w^\alpha(x) = O\left(\frac{1}{h_\alpha}\right),$$

et, par conséquent,

$$\Delta = 4 \max_{\alpha=1,\,2} \left(\max_{0 \leqslant x_\beta \leqslant l_\beta} (c_\alpha(x_\beta) + \sqrt{b_\alpha(x_\beta)})^2 \right) = O\left(\frac{1}{|h|^2}\right).$$

Donc, dans le cas de la méthode triangulaire alternée modifiée appliquée à la résolution du troisième problème aux limites (32) le nombre d'itérations dépend de celui des nœuds de la même façon que dans le cas du premier problème aux limites.

Le procédé, décrit plus haut, de réduction au problème de Dirichlet peut, apparemment, être aussi utilisé pour le cas quand à chaque côté du rectangle est imposée une des conditions aux limites, de première, de seconde ou de troisième espèce.

5. Le problème discret de Dirichlet pour équation à dérivées mixtes. Supposons que dans le rectangle $\overline{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha,\ \alpha = 1,\ 2\}$ à frontière Γ, il s'agit de trouver la solution du problème de Dirichlet pour l'équation elliptique à dérivées mixtes

$$Lu = \sum_{\alpha,\,\beta=1}^{2} \frac{2}{\partial x_\alpha}\left(k_{\alpha\beta}(x)\frac{\partial u}{\partial x_\beta}\right) = -\varphi(x), \quad x \in G,$$
$$u(x) = g(x), \quad x \in \Gamma, \quad k_{\alpha\beta}(x) = k_{\beta\alpha}(x). \tag{37}$$

Veillons à ce que soient remplies les conditions

$$k_{\alpha\alpha}(x) \geqslant c > 0, \quad k_{12}^2(x) \leqslant \rho^2 k_{11}(x)\, k_{22}(x),\ x \in \overline{G},$$
$$0 \leqslant \rho < 1. \tag{38}$$

Notons que les conditions (38) garantissent l'ellipticité régulière de l'équation (37). En effet, examinons pour un x fixé appartenant à G le problème aux valeurs propres pour un faisceau de matrices

$$\left\| \begin{matrix} k_{11} & k_{12} \\ k_{12} & k_{22} \end{matrix} \right\| - \lambda \left\| \begin{matrix} k_{11} & 0 \\ 0 & k_{22} \end{matrix} \right\| = 0.$$

Pour λ on a une équation quadratique $(1-\lambda)^2 k_{11} k_{22} - k_{12}^2 = 0$. De là il vient

$$|1-\lambda| = \frac{|k_{12}|}{\sqrt{k_{11}k_{22}}} \leqslant \rho, \quad 1-\rho \leqslant \lambda \leqslant 1+\rho.$$

Par conséquent, on a l'inégalité

$$c_1 \sum_{\alpha=1}^{2} k_{\alpha\alpha}(x)\,\xi_\alpha^2 \leqslant \sum_{\alpha,\beta=1}^{2} k_{\alpha\beta}(x)\,\xi_\alpha\xi_\beta \leqslant c_2 \sum_{\alpha=1}^{2} k_{\alpha\alpha}(x)\,\xi_\alpha^2, \tag{39}$$
$$c_1 = 1-\rho, \quad c_2 = 1+\rho,$$

où $\xi = (\xi_1, \xi_2)$ est un vecteur quelconque. De là, en vertu de la condition $k_{\alpha\alpha}(x) \geqslant c > 0$, il s'ensuit l'ellipticité régulière de l'opérateur L.

Sur un maillage rectangulaire $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leqslant \leqslant i \leqslant N_1, 0 \leqslant j \leqslant N_2, h_\alpha N_\alpha = l_\alpha, \alpha = 1, 2\}$ faisons correspondre au problème (37), (38) le problème discret de Dirichlet

$$\Lambda y = \frac{1}{2} \sum_{\alpha, \beta=1}^{2} [(k_{\alpha\beta} y_{\bar{x}_\beta})_{x_\alpha} + (k_{\alpha\beta} y_{x_\beta})_{\bar{x}_\alpha}] = -\varphi(x), \quad x \in \omega,$$
$$y(x) = g(x), \quad x \in \gamma. \tag{40}$$

Dans l'espace H des fonctions de mailles données sur ω avec produit scalaire

$$(u, v) = \sum_{x \in \omega} u(x) v(x) h_1 h_2$$

définissons l'opérateur A de la façon suivante : $Ay = -\Lambda \mathring{y}$, $y \in H$ et $y(x) = \mathring{y}(x)$ pour $x \in \omega$, $\mathring{y}(x) = 0$ pour $x \in \gamma$, de même que l'opérateur R : $Ry = -\mathscr{R}\mathring{y}$, où

$$\mathscr{R} y = \sum_{\alpha=1}^{2} (a_\alpha y_{\bar{x}_\alpha})_{x_\alpha}, \quad x \in \omega, \quad a_\alpha(x) = \frac{k_{\alpha\alpha} + k_{\alpha\alpha}^{-1}}{2}.$$

Dans ce cas le problème (40) peut être écrit sous la forme de l'équation (3), où $f(x)$ ne diffère de $\varphi(x)$ que dans les nœuds frontières. Comme $k_{\alpha\beta}(x) = k_{\beta\alpha}(x)$, les opérateurs A et R sont autoadjoints. Montrons qu'on a les inégalités

$$c_1 R \leqslant A \leqslant c_2 R, \quad c_1 = 1 - \rho, \quad c_2 = 1 + \rho, \tag{41}$$

où ρ est défini dans (38). De fait, des formules discrètes de Green il vient

$$(Ay, y) = -(\Lambda \mathring{y}, \mathring{y}) = \sum_{\alpha, \beta=1}^{2} \frac{1}{2} [(k_{\alpha\beta} \mathring{y}_{\bar{x}_\beta}, \mathring{y}_{\bar{x}_\alpha})_\alpha + {}_\alpha(k_{\alpha\beta} \mathring{y}_{x_\beta}, \mathring{y}_{x_\alpha})],$$

où le produit scalaire (u, v) est défini au point 1, § 2, tandis que

$${}_\alpha(u, v) = \sum_{x_\alpha=0}^{l_\alpha - h_\alpha} \sum_{x_\beta=h_\beta}^{l_\beta - h_\beta} u(x) v(x) h_1 h_2, \quad \beta = 3 - \alpha, \quad \alpha = 1, 2.$$

Notons que si l'une des fonctions $u(x)$ ou $v(x)$ s'annule pour $x_\beta = 0$ (ou pour $x_\beta = l_\beta$), on obtient alors

$${}_\alpha(u, v) = [u, v) = \sum_{x_1=0}^{l_1 - h_1} \sum_{x_2=0}^{l_2 - h_2} u(x) v(x) h_1 h_2,$$

$$(u, v)_\alpha = (u, v] = \sum_{x_1=h_1}^{l_1} \sum_{x_2=h_2}^{l_2} u(x)\, v(x)\, h_1 h_2, \quad \alpha = 1, 2.$$

Alors immédiatement on a

$$(Ay, y) = \sum_{\alpha, \beta=1}^{2} \frac{1}{2} \{(k_{\alpha\beta} \overset{\circ}{y}_{\bar{x}_\beta}, \overset{\bullet}{y}_{\bar{x}_\alpha}] + [k_{\alpha\beta} \overset{\circ}{y}_{x_\beta}, \overset{\bullet}{y}_{x_\alpha})\}. \qquad (42)$$

Ensuite, vu que $\mathscr{R}y$ peut être écrit sous la forme

$$\mathscr{R}y = \frac{1}{2} \sum_{\alpha=1}^{2} \{(k_{\alpha\alpha} y_{\bar{x}_\alpha})_{x_\alpha} + (k_{\alpha\alpha} y_{x_\alpha})_{\bar{x}_\alpha}\},$$

on a

$$(Ry, y) = -(\mathscr{R}\overset{\bullet}{y}, \overset{\circ}{y}) = \sum_{\alpha=1}^{2} \frac{1}{2} \{(k_{\alpha\alpha} \overset{\circ}{y}_{\bar{x}_\alpha}, \overset{\circ}{y}_{\bar{x}_\alpha})_\alpha + {}_\alpha(k_{\alpha\alpha} \overset{\bullet}{y}_{x_\alpha}, \overset{\bullet}{y}_{x_\alpha})\} =$$

$$= \sum_{\alpha=1}^{2} \frac{1}{2} \{(k_{\alpha\alpha} \overset{\circ}{y}{}^2_{\bar{x}_\alpha}, 1] + [k_{\alpha\alpha} \overset{\bullet}{y}{}^2_{x_\alpha}, 1)\}. \qquad (43)$$

De (42), (43) et des inégalités (39), on obtient

$$c_1 (\sum_{\alpha=1}^{2} k_{\alpha\alpha} \overset{\circ}{y}{}^2_{\bar{x}_\alpha}, 1] \leqslant (\sum_{\alpha, \beta=1}^{2} k_{\alpha\beta} \overset{\circ}{y}_{\bar{x}_\beta} \overset{\circ}{y}_{\bar{x}_\alpha}, 1] \leqslant c_2 (\sum_{\alpha=1}^{2} k_{\alpha\alpha} \overset{\bullet}{y}{}^2_{\bar{x}_\alpha}, 1]$$

et de façon analogue

$$c_1 [\sum_{\alpha=1}^{2} k_{\alpha\alpha} \overset{\circ}{y}{}^2_{x_\alpha}, 1) \leqslant [\sum_{\alpha, \beta=1}^{2} k_{\alpha\beta} \overset{\bullet}{y}_{x_\beta} \overset{\circ}{y}_{x_\alpha}, 1) \leqslant c_2 [\sum_{\alpha=1}^{2} k_{\alpha\alpha} \overset{\bullet}{y}{}^2_{x_\alpha}, 1),$$

et, par conséquent, les estimations (41) sont démontrées.

On peut donc utiliser l'opérateur R, défini plus haut, en qualité de régulateur dans la méthode triangulaire alternée

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \ldots,$$
$$B = (\mathscr{D} + \omega R_1) \mathscr{D}^{-1} (\mathscr{D} + \omega R_2) \quad R_1 = R_1^*, \quad R_1 + R_2 = R,$$

où les opérateurs R_1, R_2 et $\mathscr{D}$ sont définis au point 2, § 2. On y a également obtenu les constantes δ et Δ des inégalités $\delta\mathscr{D} \leqslant R$, $R_1 \mathscr{D}^{-1} R_2 < \frac{\Delta}{4} R$, $\delta > 0$. Avec l'application du théorème 2 on achève la construction de la méthode triangulaire alternée pour le problème de différences (40).

§ 3. Méthode triangulaire alternée de résolution des équations elliptiques dans un domaine arbitraire

1. Position du problème de différences. Construisons la méthode triangulaire alternée modifiée pour la résolution du problème de Dirichlet dans un domaine borné arbitraire $\bar{G}$ à frontière Γ au cas d'une équation elliptique à coefficients variables

$$\sum_{\alpha=1}^{2} \frac{\partial}{\partial x_\alpha}\left(k_\alpha(x)\frac{\partial u}{\partial x_\alpha}\right) = -\varphi(x), \quad x \in G, \tag{1}$$
$$u(x) = g(x), \quad x \in \Gamma, \quad k_\alpha(x) \geqslant c_1 > 0, \quad \alpha = 1, 2.$$

Posons que la frontière Γ est suffisamment lisse. En outre, pour simplifier l'exposé, admettons que l'intersection du domaine avec la droite passant par tout point $x \in G$ parallèlement à l'axe des coordonnées Ox_α, $\alpha = 1, 2$, constitue un seul intervalle.

Construisons dans le domaine $\bar{G}$ un maillage irrégulier $\bar{\omega}$ de la façon suivante. Traçons une famille de droites $x_\alpha = x_\alpha(i_\alpha)$, $i_\alpha = 0, \pm 1, \pm 2, \ldots, \alpha = 1, 2$. Dans ce cas les points $x_i = (x_1(i_1), x_2(i_2))$, $i = (i_1, i_2)$ forment dans le plan un maillage principal. Appelons le point x_i du maillage appartenant à G nœud intérieur du maillage $\bar{\omega}$. Désignons comme d'habitude l'ensemble de tous les nœuds intérieurs par ω.

L'intersection de toute droite tracée par le point $x_i \in \omega$ parallèlement à l'axe Ox_α avec le domaine G constitue l'intervalle $\Delta_\alpha(x_i)$. Les extrémités de cet intervalle seront appelées nœuds frontières en direction de x_α. Notons l'ensemble de tous les nœuds frontières suivant x_α par γ_α. La frontière du maillage $\bar{\omega}$ est $\gamma = \gamma_1 \cup \gamma_2$, de sorte que $\bar{\omega} = \omega \cup \gamma$. Le malliage $\bar{\omega}$ est ainsi construit.

Introduisons une série de notations. Désignons par $\omega_\alpha(x_\beta)$, $\beta = 3 - \alpha$, $\alpha = 1, 2$ l'ensemble des nœuds du maillage ω se disposant sur l'intervalle Δ_α; $\omega_\alpha^+(x_\beta)$ l'ensemble comprenant $\omega_\alpha(x_\beta)$ et l'extrémité droite Δ_α de l'intervalle; $\bar{\omega}_\alpha(x_\beta)$ est composé de $\omega_\alpha(x_\beta)$ et des extrémités de l'intervalle Δ_α.

Désignons par $x^{(+1_\alpha)}$ et $x^{(-1_\alpha)}$ les nœuds voisins de $x \in \omega_\alpha(x_\beta)$ à droite et à gauche et appartenant à $\bar{\omega}_\alpha(x_\beta)$. Notons que si, par exemple, $x^{(+1_\alpha)} \in \gamma_\alpha$, ce nœud peut ne pas se confondre avec le nœud du maillage principal.

Définissons $h_\alpha^+(x) = x^{(+1_\alpha)} - x$, $h_\alpha^-(x) = x - x^{(-1_\alpha)}$, $x \in \omega_\alpha$, $x^{(\pm 1_\alpha)} \in \bar{\omega}_\alpha$. Définissons également dans tous les nœuds intérieurs du maillage ω les pas moyens $\hbar_\alpha(x_\alpha) = 0{,}5\,(x_\alpha(i_\alpha + 1) - x_\alpha(i_\alpha - 1))$ comme la distance séparant les droites correspondantes du maillage principal.

Mettons en correspondance le problème (1) associé au maillage $\bar{\omega}$ et le problème de différences

$$\Lambda y = \sum_{\alpha=1}^{2} (a_\alpha y_{\bar{x}_\alpha})_{\hat{x}_\alpha} = -\varphi(x), \; x \in \omega, \qquad y(x) = g(x), \quad x \in \gamma. \tag{2}$$

On a utilisé ici les notations suivantes:

$$(a_\alpha x_{\bar{x}_\alpha})_{\hat{x}_\alpha} = \frac{1}{\hbar_\alpha} (a_\alpha^{+1} y_{x_\alpha} - a_\alpha y_{\bar{x}_\alpha}), \; a_\alpha^{+1} = a_\alpha (x^{(+1_\alpha)}),$$

$$y_{x_\alpha} = \frac{1}{h_\alpha^+} (y(x^{(+1_\alpha)}) - y(x)), \; y_{\bar{x}_\alpha} = \frac{1}{h_\alpha^-} (y(x) - y(x^{(-1_\alpha)})).$$

Les coefficients $a_\alpha(x)$ et $\varphi(x)$ sont choisis de façon que le schéma (2) sur le maillage régulier possède le second ordre local d'approximation.

Introduisons maintenant l'espace H des fonctions de mailles associées à ω avec produit scalaire $(u, v) = \sum_{x \in \omega} u(x) v(x) \hbar_1(x_1) \hbar_2(x_2)$. L'opérateur A sera défini comme habituellement: $Ay = -\Lambda \mathring{y}$, $y \in H$ et $y(x) = \mathring{y}(x)$ pour $x \in \omega$, $\mathring{y}(x) = 0$ pour $x \in \gamma$. Alors le problème de différences (2) s'écrira sous la forme de l'équation

$$Au = f, \tag{3}$$

où $f(x)$ ne diffère de $\varphi(x)$ que dans les nœuds frontières.

2. Construction de la méthode triangulaire alternée. Etudions pour l'équation (3) la méthode triangulaire alternée modifiée

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \; k = 0, 1, \ldots, \qquad B = (\mathscr{D} + \omega R_1) \mathscr{D}^{-1} (\mathscr{D} + \omega R_2), \; R_1 = R_1^*, \quad R_1 + R_2 = A. \tag{4}$$

Définissons les opérateurs R_1, R_2 et $\mathscr{D}$. Comme dans le cas du rectangle, choisissons le plus simple des opérateurs $\mathscr{D}$:

$$\mathscr{D} y = d(x) y, \quad d(x) > 0 \text{ pour } x \in \omega, \tag{5}$$

et posons $R_\alpha y = -\mathscr{R}_\alpha \mathring{y}$, $y \in H$ et $y(x) = \mathring{y}(x)$, $x \in \omega$, où

$$\mathscr{R}_1 y = -\sum_{\alpha=1}^{2} \left[\frac{a_\alpha}{\hbar_\alpha} y_{\bar{x}_\alpha} + \frac{1}{2\hbar_\alpha} \left(\frac{a_\alpha^{+1}}{h_\alpha^+} - \frac{a_\alpha}{h_\alpha^-} \right) y \right], \qquad \mathscr{R}_2 y = \sum_{\alpha=1}^{2} \left[\frac{a_\alpha^{+1}}{\hbar_\alpha} y_{x_\alpha} + \frac{1}{2\hbar_\alpha} \left(\frac{a_\alpha^{+1}}{h_\alpha^+} - \frac{a_\alpha}{h_\alpha^-} \right) y \right]. \tag{6}$$

Vu que $\mathscr{R}_1 + \mathscr{R}_2 = \Lambda$, on aboutit à ce que $R_1 + R_2 = A$.

Montrons que les opérateurs R_1 et R_2 sont autoadjoints. Introduisons d'abord les notations qu'on utilisera par la suite. Définissons

les sommes suivantes:

$$(u, v)_{\omega_\alpha(x_\beta)} = \sum_{x_\alpha \in \omega_\alpha(x_\beta)} u(x)\, v(x)\, \hbar_\alpha(x_\alpha),$$

$$(u, v)_{\omega_\alpha^+(x_\beta)} = \sum_{x_\alpha \in \omega_\alpha^+(x_\beta)} u(x)\, v(x)\, h_\alpha^-(x),$$

$$(u, v)_\alpha = ((u, v)_{\omega_\alpha^+(x_\beta)},\ 1)_{\omega_\beta(x_\alpha)} = \sum_{x_\beta \in \omega_\beta} \sum_{x_\alpha \in \omega_\alpha^+} u(x) v(x\, h_\alpha^-(x)\, \hbar_\beta(x_\beta),$$

$$\beta = 3 - \alpha, \quad \alpha = 1, 2.$$

En utilisant les notations introduites, on peut écrire le produit scalaire dans H de la façon suivante:

$$(u, v) = ((u, v)_{\omega_1(x_2)},\ 1)_{\omega_2(x_1)} = ((u, v)_{\omega_2(x_1)},\ 1)_{\omega_1(x_2)}. \tag{7}$$

Démontrons d'abord une proposition auxiliaire. Soient y_i et v_i des fonctions de mailles données pour $0 \leqslant i \leqslant N$ avec $y_0 = y_N = 0$ et $v_0 = v_N = 0$. Soit u_i une fonction de maille donnée pour $1 \leqslant i \leqslant N$. On a alors l'égalité

$$\sum_{i=1}^{N-1} (u_{i+1} - u_i)\, y_i v_i = -\sum_{i=1}^{N-1} (v_{i+1} - v_i)\, u_{i+1} v_i - \sum_{i=1}^{N-1} (y_i + y_{i-1})\, u_i v_i. \tag{8}$$

On a en effet:

$$\sum_{i=1}^{N-1} [(u_{i+1} - u_i)\, y_i v_i + (v_{i+1} - v_i)\, u_{i+1} y_i + (y_i - y_{i-1}) u_i v_i] =$$
$$= \sum_{i=1}^{N-1} (u_{i+1} v_{i+1} y_i - u_i v_i y_{i-1}) = u_N v_N y_{N-1} - u_1 v_1 y_0 = 0.$$

La proposition (8) est démontrée. En utilisant (8), on montre facilement que pour les fonctions $\overset{\circ}{y}(x)$ et $\overset{\circ}{v}(x)$ données sur $\overline{\omega}$ et s'annulant sur γ on a l'égalité

$$(u_{x_\alpha} \overset{\circ}{y}, \overset{\circ}{v})_{\omega_\alpha} = -\left(\overset{\circ}{y}, \frac{h_\alpha^+}{h_\alpha} u^{+1_\alpha} \overset{\circ}{v}_{x_\alpha}\right)_{\omega_\alpha} - \left(\frac{h_\alpha^-}{h_\alpha} u \overset{\circ}{y}_{\bar{x}_\alpha}, \overset{\circ}{v}\right)_{\omega_\alpha}. \tag{9}$$

On a utilisé ici les notations

$$u^{+1_\alpha} = -u\,(x^{(+1_\alpha)}), \qquad u_{\hat{x}_\alpha} = \frac{u^{+1_\alpha} - u}{h_\alpha(x_\alpha)}.$$

En portant dans (9) l'expression $u(x) = a_\alpha(x)/h_\alpha^-(x)$ et compte tenu de l'égalité $h_\alpha^-(x^{(+1_\alpha)}) = h_\alpha^+(x)$, il vient

$$\left(\frac{1}{h_\alpha}\left(\frac{a_\alpha^{+1}}{h_\alpha^+} - \frac{a_\alpha}{h_\alpha^-}\right) \overset{\circ}{y}, \overset{\circ}{v}\right)_{\omega_\alpha} = -\left(\overset{\circ}{y}, \frac{a_\alpha^{+1}}{\hbar_\alpha} \overset{\circ}{v}_{x_\alpha}\right)_{\omega_\alpha} - \left(\frac{a_\alpha}{h_\alpha} \overset{\circ}{y}_{\bar{x}_\alpha}, \overset{\circ}{v}\right)_\alpha.$$

En multipliant cette relation par $\hbar_\beta(x_\beta)$, en sommant en ω_β et en tenant compte de (7), on obtient

$$-\left(\frac{a_\alpha}{h_\alpha}\mathring{y}_{\bar{x}_\alpha},\ \mathring{v}\right)=\left(\frac{a_\alpha^{+1}}{\hbar_\alpha}\mathring{v}_{x_\alpha},\ \mathring{y}\right)+\left(\frac{1}{\hbar_\alpha}\left(\frac{a_\alpha^{+1}}{h_\alpha^+}-\frac{a_\alpha}{h_\alpha^-}\right)\mathring{y},\ \mathring{v}\right). \quad (10)$$

Démontrons à présent que les opérateurs R_1 et R_2 sont autoconjugués. De (6) et (10), il vient

$$(R_1y,\ v)=-(\mathcal{R}_1\mathring{y},\ \mathring{v})=$$
$$=\sum_{\alpha=1}^{2}\left[\left(\frac{a_\alpha}{h_\alpha}\mathring{y}_{\bar{x}_\alpha},\ \mathring{v}\right)+\left(\frac{1}{2\hbar_\alpha}\left(\frac{a_\alpha^{+1}}{h_\alpha^+}-\frac{a_\alpha}{h_\alpha^-}\right)\mathring{x},\ \mathring{v}\right)\right]=$$
$$=-\sum_{\alpha=1}^{2}\left[\left(\mathring{y},\ \frac{a_\alpha^{+1}}{\hbar_\alpha}\mathring{v}_{x_\alpha}\right)+\left(\frac{1}{2\hbar_\alpha}\left(\frac{a_\alpha^{+1}}{h_\alpha^+}-\frac{a_\alpha}{h_\alpha^-}\right)\mathring{y},\ \mathring{v}\right)\right]=$$
$$=-(\mathring{y},\ \mathcal{R}_2\mathring{v})=(y,\ R_2v).$$

La proposition est démontrée. D'ailleurs de là s'ensuit la conclusion que l'opérateur A est autoadjoint.

Il ne reste qu'à construire la fonction $d(x)$ déterminant l'opérateur $\mathcal{D}$, à rechercher les constantes δ et Δ des inégalités

$$\delta\mathcal{D}\leqslant A,\ R_1\mathcal{D}^{-1}R_2\leqslant\frac{\Delta}{4}A,\ \delta>0 \quad (11)$$

et à se servir du théorème 1. Tout cela, on le réalisera comme au point 2, § 2, où on a étudié le problème de Dirichlet pour l'équation elliptique dans un rectangle sur un maillage régulier.

Notons d'abord qu'en vertu des formules de différences de Green on a l'égalité

$$(Ay,\ y)=\sum_{\alpha=1}^{2}(a_\alpha\dot{y}_{\bar{x}_\alpha}^2,\ 1)_\alpha,\ \dot{y}(x)=0,\ x\in\gamma.$$

Ensuite, de (5) et (6) on trouve

$$(R_1\mathcal{D}^{-1}R_2y,\ y)=(\mathcal{D}^{-1}\mathcal{R}_2\dot{y},\ \mathcal{R}_2\dot{y})=$$
$$=\left(\frac{1}{d}\sum_{\alpha=1}^{2}\left[\frac{a_\alpha^{+1}}{\hbar_\alpha}\mathring{y}_{x_\alpha}+\frac{1}{2\hbar_\alpha}\left(\frac{a_\alpha^{+1}}{h_\alpha^+}-\frac{a_\alpha}{h_\alpha^-}\right)y\right]^2,\ 1\right).$$

Utilisons à présent le lemme 2 en posant

$$p_\alpha=\frac{a_\alpha^{+1}}{\hbar_\alpha},\quad q_\alpha=\frac{h_\alpha^{+1}}{2\hbar_\alpha}\left(\frac{a_\alpha^{+1}}{h_\alpha^+}-\frac{a_\alpha}{h_\alpha^-}\right),$$
$$u_\alpha=\mathring{y}_{x_\alpha},\quad v_\alpha=\frac{1}{h_\alpha^+}\mathring{y},\quad \alpha=1,\ 2.$$

Finalement, on obtient l'inégalité

$$(R_1\mathscr{D}^{-1}R_2y,\ y)\leqslant\Big(\frac{(1+\varepsilon)}{d\hbar_1^2}\Big[a_1^{+1}+\frac{\varkappa_1h_1^+}{2}\Big|\frac{a_1^{+1}}{h_1^+}-\frac{a_1}{h_1^-}\Big|\Big]\times$$
$$\times\Big[a_1^{+1}\overset{\circ}{y}{}^2_{\bar{x}_1}+\frac{1}{2\varkappa_1h_1^+}\Big|\frac{a_1^{+1}}{h_1^+}-\frac{a_1}{h_1^-}\Big|\overset{\circ}{y}{}^2\Big],\ 1\Big)+$$
$$+\Big(\frac{(1+\varepsilon)}{d\varepsilon\hbar_2^2}\Big[a_2^{+1}+\frac{\varkappa_2h_2^+}{2}\Big|\frac{a_2^{+1}}{h_2^+}-\frac{a_2}{h_2^-}\Big|\Big]\times$$
$$\times\Big[a_2^{+1}\overset{\circ}{y}{}^2_{\bar{x}_2}+\frac{1}{2\varkappa_2h_2^+}\Big|\frac{a_2^{+1}}{h_2^+}-\frac{a_2}{h_2^-}\Big|\overset{\circ}{y}{}^2\Big)\Big],1.$$

Posons ici

$$\varepsilon=\varepsilon(x)=\frac{a_2^{+1}+0{,}5\varkappa_2h_2^+\left|\dfrac{a_2^{+1}}{h_2^+}-\dfrac{a_2}{h_2^-}\right|}{a_1^{+1}+0{,}5\varkappa_1h_1^+\left|\dfrac{a_1^{+1}}{h_1^+}-\dfrac{a_1}{h_1^-}\right|}\,\frac{\hbar_1h_1^+}{\hbar_2h_2^+}\,\frac{\theta_2}{\theta_1}$$

et déterminons $d(x)$ de la façon suivante:

$$d(x)=\sum_{\alpha=1}^{2}\Big(a_\alpha^{+1}+\frac{\varkappa_\alpha h_\alpha^+}{2}\Big|\frac{a_\alpha^{+1}}{h_\alpha^+}-\frac{a_\alpha}{h_\alpha^-}\Big|\Big)\frac{\theta_\alpha}{\hbar_\alpha h_\alpha^+},\quad x\in\omega.$$

On suppose ici que $\varkappa_\alpha=\varkappa_\alpha(x_\beta)>0$, $\theta_\alpha=\theta_\alpha(x_\beta)>0$, $\beta=3-\alpha$, $\alpha=1,\ 2$. Finalement, on obtient l'inégalité

$$(R_1\mathscr{D}^{-1}R_2y,\ y)\leqslant$$
$$\leqslant\sum_{\alpha=1}^{2}\Big(\frac{a_\alpha^{+1}h_\alpha^+}{\theta_\alpha\hbar_\alpha}\overset{\circ}{y}{}^2_{\bar{x}_\alpha},\ 1\Big)+\sum_{\alpha=1}^{2}\Big(\frac{1}{2\hbar_\alpha\theta_\alpha\varkappa_\alpha}\Big|\frac{a_\alpha^{+1}}{h_\alpha^+}-\frac{a_\alpha}{h_\alpha^-}\Big|\overset{\circ}{y}{}^2,\ 1\Big).$$

Etant donné que θ_α ne dépend pas de x_α, il s'ensuit que

$$\Big(\frac{a_\alpha^{+1}}{\theta_\alpha}\,\frac{h_\alpha^+}{\hbar_\alpha}\overset{\circ}{y}{}^2_{\bar{x}_\alpha},\ 1\Big)_{\omega_\alpha}=\frac{1}{\theta_\alpha}\sum_{x_\alpha\in\omega_\alpha}a_\alpha^{+1}h_\alpha^+\overset{\circ}{y}{}^2_{\bar{x}_\alpha}\leqslant\frac{1}{\theta_\alpha}\sum_{x_\alpha\in\omega_\alpha^+}a_\alpha h_\alpha^-\overset{\circ}{y}{}^2_{\bar{x}_\alpha}.$$

Par conséquent,

$$\Big(\frac{a_\alpha^{+1}}{\theta_\alpha}\,\frac{h_\alpha^+}{\hbar_\alpha}\overset{\circ}{y}{}^2_{\bar{x}_\alpha},\ 1\Big)\leqslant\Big(\frac{a_\alpha}{\theta_\alpha}\overset{\circ}{y}{}^2_{\bar{x}_\alpha},\ 1\Big)_\alpha,$$

et on a donc finalement

$$(R_1\mathscr{D}^{-1}R_2y,\ y)\leqslant$$
$$\leqslant\sum_{\alpha=1}^{2}\Big(\frac{a_\alpha}{\theta_\alpha}\overset{\circ}{y}{}^2_{\bar{x}_\alpha},\ 1\Big)_\alpha+\sum_{\alpha=1}^{2}\Big(\frac{1}{2\hbar_\alpha\theta_\alpha\varkappa_\alpha}\Big|\frac{a_\alpha^{+1}}{h_\alpha^+}-\frac{a_\alpha}{h_\alpha^-}\Big|\overset{\circ}{y}{}^2,\ 1\Big).$$

Les calculs subséquents sont des analogues des transformations et estimations obtenues au point 2 du § 2. Donnons le résultat final: dans les inégalités (11)

$$\delta=1,\ \Delta=4\max_{\alpha=1,\,2}\left(\max_{x_\beta\in\omega_\beta}\left(c_\alpha(x_\beta)+\sqrt{b_\alpha(x_\beta)}\right)^2\right),\ \beta=3-\alpha,$$

où

$$b_\alpha(x_\beta)=\max_{x_\alpha\in\omega_\alpha} v^\alpha(x),\ c_\alpha(x_\beta)=\max_{x_\alpha\in\omega_\alpha} w^\alpha(x),\ x_\beta\in\omega_\beta;$$

la fonction $v^\alpha(x)$ est la solution du problème aux limites triponctuel

$$\begin{gathered}(a_\alpha v^\alpha_{\bar{x}_\alpha})_{\hat{x}_\alpha}=-\frac{a_\alpha^{+1}}{\hbar_\alpha h_\alpha^+},\quad x_\alpha\in\omega_\alpha(x_\beta),\\ v^\alpha(x)=0,\quad x_\alpha\in\gamma_\alpha,\end{gathered}\qquad(12)$$

tandis que la fonction $w^\alpha(x)$ est la solution du problème

$$\begin{gathered}(a_\alpha w^\alpha_{\bar{x}_\alpha})_{\hat{x}_\alpha}=-\frac{1}{2\hbar_\alpha}\left|\frac{a_\alpha^{+1}}{h_\alpha^+}-\frac{a_\alpha}{h_\alpha^-}\right|,\quad x_\alpha\in\omega_\alpha(x_\beta),\\ w^\alpha(x)=0,\quad x_\alpha\in\gamma_\alpha.\end{gathered}\qquad(13)$$

La fonction $d(x)$ se calcule dans ce cas suivant la formule

$$d(x)=\sum_{\alpha=1}^{2}\left(\frac{a_\alpha^{+1}}{\hbar_\alpha h_\alpha^+\sqrt{b_\alpha}}+\frac{1}{2\hbar_\alpha}\left|\frac{a_\alpha^{+1}}{h_\alpha^+}-\frac{a_\alpha}{h_\alpha^-}\right|\right)\frac{1}{c_\alpha+\sqrt{b_\alpha}},\quad x\in\omega.$$

Les paramètres d'itérations ω et $\{\tau_k\}$ se calculent suivant les formules du théorème 1. Pour obtenir y_{k+1}, on peut se servir de l'algorithme

$$(v(x)=\alpha_1(x)v^{(-1_1)}+\beta_1(x)v^{(-1_2)}+\varkappa(x)\varphi_k(x),\quad x\in\omega,$$

$$v(x)=0,\ x\in\gamma,$$

$$\overset{\circ}{y}_{k+1}(x)=\alpha_2(x)\overset{\circ}{y}{}^{(+1_1)}_{k+1}+\beta_2(x)\overset{\circ}{y}{}^{(+1_2)}_{k+1}+\varkappa(x)d(x)v(x),\ x\in\omega,$$

$$\overset{\circ}{y}_{k+1}(x)=0,\quad x\in\gamma.$$

où

$$\alpha_1=\frac{\omega_0 a_1\varkappa}{\hbar_1 h_1^-},\quad \beta_1=\frac{\omega_0 a_2\varkappa}{\hbar_2 h_2^-},\quad \alpha_2=\frac{\omega_0 a_1^{+1}\varkappa}{\hbar_1 h_1^+},\quad \beta_2=\frac{\omega_0 a_2^{+1}\varkappa}{\hbar_2 h_2^+},$$

$$\frac{1}{\varkappa}=d+\frac{1}{2}\sum_{\alpha=1}^{2}\frac{\omega_0}{\hbar_\alpha}\left(\frac{a_\alpha^{+1}}{h_\alpha^+}+\frac{a_\alpha}{h_\alpha^-}\right),\quad \varphi_k(x)=By_k-\tau_{k+1}(Ay_k-f).$$

Il faut remarquer que dans des cas analogues, quand le calcul de la valeur de By_k est très laborieux et il n'est pas possible de limiter le volume de l'information intermédiaire mémorisée, il est rationnel d'utiliser le second algorithme décrit au point 1, § 1.

3. Problème de Dirichlet pour l'équation de Poisson dans un domaine quelconque. En guise d'exemple, voyons comment la méthode construite s'applique au problème de Dirichlet pour l'équation de Poisson

$$\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}=-\varphi(x), \quad x\in G, \ u(x)=g(x), \ x\in\Gamma.$$

Admettons que le maillage est carré, c'est-à-dire que $x_\alpha = x_\alpha(i_\alpha)$, $x_\alpha(i_\alpha+1) = x_\alpha(i_\alpha)+h$, $i_\alpha = 0, \pm 1, \pm 2, \ldots$, et

$$\omega = \{x_i = (i_1h, i_2h)\in G, \ i_\alpha = 0, \pm 1, \pm 2, \ldots\}.$$

En outre, $\hbar_\alpha \equiv h$, tandis que les pas $h_\alpha^\pm(x)$ ne sont différents de h que dans les nœuds frontières du maillage $\bar\omega$.

Prenons le schéma aux différences (2) dans lequel on pose $a_\alpha(x) \equiv \equiv 1$ et $\hbar_\alpha \equiv h$. Pour appliquer la méthode triangulaire alternée construite au point 2 du § 3, il faut trouver la solution des problèmes aux limites triponctuels (12) et (13) possédant une dimension, qui, dans le cas concerné, ont la forme

$$\Lambda_\alpha v^\alpha = v^\alpha_{\bar x_\alpha \hat x_\alpha} = -\frac{1}{hh_\alpha^+}, \quad x_\alpha\in\omega_\alpha(x_\beta), \ v^\alpha(x)=0, \ x_\alpha\in\gamma_\alpha, \qquad (14)$$

$$\Lambda_\alpha w^\alpha = w^\alpha_{\bar x_\alpha \hat x_\alpha} = -\frac{1}{2h}\left|\frac{1}{h_\alpha^+}-\frac{1}{h_\alpha^-}\right|, \qquad (15)$$
$$x_\alpha\in\omega_\alpha(x_\beta), \quad w^\alpha(x)=0, \quad x_\alpha\in\gamma_\alpha.$$

Examinons l'intervalle Δ_α contenant $\omega_\alpha(x_\beta)$, et désignons par $l_\alpha(x_\beta)$ et $L_\alpha(x_\beta)$ ses extrémités gauche et droite. Quand $h_\alpha^-(x)\leqslant h$, si x est le nœud du maillage ω_α le plus proche de l_α, et $h_\alpha^+(x)\leqslant h$, si x est le nœud du maillage ω_α le plus proche de L_α. Tous les autres pas $h_\alpha^\pm$ sont égaux au pas principal de maillage h.

L'opérateur Λ_α s'écrit dans ce cas en détail sur le maillage ω_α de la façon suivante:

$$\Lambda_\alpha y = \begin{cases} \dfrac{1}{h}\left(\dfrac{y^{+1_\alpha}-y}{h}-\dfrac{y-y^{-1_\alpha}}{h_\alpha^-}\right), & x_\alpha = l_\alpha + h_\alpha^-, \\ \dfrac{1}{h^2}(y^{+1_\alpha}-2y+y^{-1_\alpha}), & l_\alpha+h_\alpha^-+h\leqslant x_\alpha\leqslant L_\alpha-h_\alpha^+-h, \\ \dfrac{1}{h}\left(\dfrac{y^{+1_\alpha}-y}{h_\alpha^+}-\dfrac{y-y^{-1_\alpha}}{h}\right), & x_\alpha = L_\alpha - h_\alpha^+. \end{cases}$$

La solution des équations (14), (15) peut être obtenue sous une forme explicite. A cette fin, portons les conditions aux limites dans les équations écrites aux points $x_\alpha = l_\alpha + h_\alpha^-$ et $x_\alpha = L_\alpha - h_\alpha^+$. Ces équations se transforment alors et deviennent biponctuelles; on peut les considérer comme des conditions aux limites nouvelles pour les équations triponctuelles à coefficients constants écrits pour $l_\alpha + h_\alpha^- + h \leqslant x_\alpha \leqslant L_\alpha - h_\alpha^+ - h$. On aura donc les problèmes suivants pour $v(x)$ et $w(x)$ (indice supérieur α est omis temporairement pour v et w):

$$\begin{gathered} v^{+1_\alpha} - 2v + v^{-1_\alpha} = -1, \quad l_\alpha + h_\alpha^- + h \leqslant x_\alpha \leqslant L_\alpha - h_\alpha^+ - h, \\ \left(1 + \frac{h}{h_\alpha^-}\right) v = v^{+1_\alpha} + 1, \quad x_\alpha = l_\alpha + h_\alpha^-, \\ \left(1 + \frac{h_\alpha^+}{h}\right) v = \frac{h_\alpha^+}{h} v^{-1_\alpha} + 1, \quad x_\alpha = L_\alpha - h_\alpha^+, \end{gathered} \tag{16}$$

$$\begin{gathered} w^{+1_\alpha} - 2w + w^{-1_\alpha} = 0, \quad l_\alpha + h_\alpha^- + h \leqslant x_\alpha \leqslant L_\alpha - h_\alpha^+ - h, \\ \left(1 + \frac{h}{h_\alpha^-}\right) w = w^{+1_\alpha} - \frac{1}{2}\left(1 - \frac{h}{h_\alpha^-}\right), \quad x_\alpha = l_\alpha + h_\alpha^-, \\ \left(1 + \frac{h_\alpha^+}{h}\right) w = \frac{h_\alpha^+}{h} w^{-1_\alpha} + \frac{1}{2}\left(1 - \frac{h_\alpha^+}{h}\right), \quad x_\alpha = L_\alpha - h_\alpha^+. \end{gathered} \tag{17}$$

En utilisant les méthodes de résolution des équations à coefficients constants exposées au § 4 du ch. I, on aboutit à la forme explicite de la solution des problèmes aux limites (16) et (17):

$$v_\alpha(x) = \frac{1}{2h^2}\left[(x_\alpha - l_\alpha)\left(L_\alpha - x_\alpha + \frac{2h^2 - (h_\alpha^+ + h_\alpha^-)(h_\alpha^+ + h - h_\alpha^-)}{L_\alpha - l_\alpha}\right) + \right. \\ \left. + h_\alpha^-(h - h_\alpha^-)\right],$$

$$w^\alpha(x) = \frac{1}{2} - \frac{h_\alpha^-(L_\alpha - x_\alpha) + h_\alpha^+(x_\alpha - l_\alpha)}{2h(L_\alpha - l_\alpha)}$$

pour $l_\alpha + h_\alpha^- \leqslant x_\alpha \leqslant L_\alpha - h_\alpha^+$. Comme $h_\alpha^\pm \leqslant h$, on a

$$(h_\alpha^+ + h_\alpha^-)(h_\alpha^+ + h - h_\alpha^-) \geqslant h_\alpha^-(h - h_\alpha^-),$$

donc

$$v^\alpha(x) \leqslant \frac{1}{2h^2}(x_\alpha - l_\alpha)(L_\alpha - x_\alpha) + 1 \leqslant \frac{1}{2h^2}\left(\frac{L_\alpha - l_\alpha}{2}\right)^2 + 1,$$

$$w^\alpha(x) \geqslant \frac{1}{2}, \quad \alpha = 1, 2.$$

Par conséquent,

$$b_\alpha(x_\beta) = \max_{x_\alpha \in \omega_\alpha} v^\alpha(x) \leqslant \frac{1}{2h^2}\left(\frac{L_\alpha - l_\alpha}{2}\right)^2 + 1,$$

$$c_\alpha(x_\beta) = \max_{x_\alpha \in \omega_\alpha} w^\alpha(x) \leqslant \frac{1}{2}.$$

Par suite, $\Delta = O(l_0^2/h^2)$, où l_0 est le diamètre du domaine G. Aussi, en vertu du théorème 1, voit-on se vérifier l'estimation

$$n \geqslant n_0(\varepsilon) = \frac{\ln(2/\varepsilon)}{2\sqrt{2}\sqrt[4]{\eta}} + \frac{\ln(2/\varepsilon)}{2\sqrt{2}\sqrt[4]{2}\sqrt{h/l_0}} \approx 0{,}298\sqrt{N}\ln\frac{2}{\varepsilon}, \quad (18)$$

où N est le nombre maximal de nœuds suivant la direction x_1 ou x_2. Donc le nombre d'itérations pour l'exemple modèle étudié dépend du pas principal h du maillage et ne dépend pas des pas aux nœuds frontières du maillage $\bar{\omega}$.

Comparons l'estimation (18) avec l'estimation du nombre d'itérations pour le cas du problème de Dirichlet dans un carré de côté l_0 et avec le nombre de nœuds N suivant chaque direction du maillage carré $\bar{\omega}$. L'estimation correspondante du nombre d'itérations a été obtenue au point 4 du § 1, elle est de la forme

$$n \geqslant n_0(\varepsilon) = 0{,}28\sqrt{N}\ln(2/\varepsilon).$$

Il s'ensuit que pour un domaine arbitraire G le nombre d'itérations de la méthode triangulaire alternée modifiée est le même que celui obtenu pour le même problème de Dirichlet pour l'équation de Poisson dans un carré dont le côté est égal au diamètre du domaine G.

R e m a r q u e 1. Le procédé, exposé ici, de construction de la méthode triangulaire alternée peut évidemment être utilisé dans le cas où il s'agit de résoudre l'équation elliptique dans un rectangle mais sur un maillage irrégulier.

R e m a r q u e 2. La construction de la méthode pour le cas de l'équation à dérivées mixtes peut être réalisée en recourant au choix du régulateur R de la façon analogue à ce qu'il a été déjà réalisé au point 5 du § 2.

CHAPITRE XI

MÉTHODE DES DIRECTIONS ALTERNÉES

On étudie dans ce chapitre les méthodes itératives spéciales proposées pour la résolution des équations de mailles elliptiques du type $Au = f$, dont l'opérateur A possède une structure déterminée. On expose au § 1 la méthode des directions alternées au cas de commutativité; on construit le jeu optimal des paramètres. Au § 2 la méthode est illustrée d'exemples de résolution des problèmes aux limites pour des équations elliptiques à variables séparables. Le § 3 est consacré à la méthode des directions alternées au cas de non-commutativité.

§ 1. Méthode des directions alternées au cas de commutativité

1. Schéma itératif de la méthode. On a étudié au ch. X la méthode itérative triangulaire alternée universelle dont l'opérateur B était choisi en tenant compte du développement de l'opérateur A en une somme de deux opérateurs mutuellement adjoints. Le plus souvent on recourt à un développement de l'opérateur A en une somme d'opérateurs triangulaires, B étant un produit d'opérateurs triangulaires dépendant d'un paramètre d'itération supplémentaire. La prise en compte de la structure de l'opérateur B permet de choisir de façon optimale les paramètres d'itération et de bâtir la méthode convergeant beaucoup plus vite que la méthode explicite. Appliquée à la résolution des équations de mailles elliptiques, cette méthode s'avère aussi plus économique, vu que la mise en œuvre d'une itération exige un nombre d'opérations arithmétiques proportionnel à celui d'inconnues dans le problème.

Comme on le sait, les opérateurs A, associés aux équations de mailles elliptiques, possèdent une structure spécifique. Aussi, lors du choix des opérateurs B, est-il naturel de tenter d'utiliser cette particularité de l'opérateur A dans les schémas itératifs implicites. Ces méthodes itératives ne seront évidemment pas des méthodes universelles, mais le rétrécissement de la classe des problèmes initiaux par l'exigence d'une structure déterminée de l'opérateur A autorise de construire des méthodes itératives à convergence rapide destinées, notamment, à la résolution des équations de mailles.

Le présent chapitre sera consacré à l'étude de la méthode spéciale dite *méthode itérative des directions alternées.* On fournira d'abord la description de la méthode en sa forme opératorielle, ensuite, en recourant à des exemples, on montrera comment cette méthode s'applique à la recherche de la solution approchée des différentes équations de mailles elliptiques.

Commençons la description de la méthode par le schéma itératif. Supposons qu'il s'agit de trouver la solution de l'équation opératorielle linéaire

$$Au = f \tag{1}$$

à opérateur A non dégénéré, donné dans l'espace hilbertien H. Supposons que l'opérateur A est représenté sous forme de somme de deux opérateurs A_1 et A_2, autrement dit $A = A_1 + A_2$. Pour obtenir la solution approchée de l'équation (1), prenons le schéma itératif implicite à deux couches

$$B_{k+1}\frac{y_{k+1}-y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0,\ 1,\ \ldots,\ y_0 \in H, \tag{2}$$

$$B_k = (\omega_k^{(1)}E + A_1)(\omega_k^{(2)}E + A_2), \quad \tau_k = \omega_k^{(1)} + \omega_k^{(2)}, \tag{3}$$

dans lequel l'opérateur B_{k+1} de la couche supérieure est fonction du numéro d'itération k. $\omega_k^{(1)}$ et $\omega_k^{(2)}$ sont ici des paramètres d'itération qui dépendent également du numéro d'itération k et qu'on doit déterminer.

Arrêtons-nous d'abord sur les modes de recherche de y_{k+1} avec y_k donné. Un des algorithmes possibles de mise en œuvre du schéma (2) est le suivant:

$$\begin{gathered}(\omega_{k+1}^{(1)}E + A_1)\, y_{k+1/2} = (\omega_{k+1}^{(1)}E - A_2)\, y_k + f,\\ (\omega_{k+1}^{(2)}E + A_2)\, y_{k+1} = (\omega_{k+1}^{(2)}E - A_1)\, y_{k+1/2} + f,\quad k = 0,\ 1,\ \ldots,\end{gathered} \tag{4}$$

où $y_{k+1/2}$ est une approximation itérative intermédiaire.

Montrons que le système (4) est algébriquement équivalent au schéma (2). Pour cela, excluons $y_{k+1/2}$ de (4). Compte tenu de ce que $A = A_1 + A_2$, récrivons (4) sous la forme suivante:

$$\begin{gathered}(\omega_{k+1}^{(1)}E + A_1)(y_{k+1/2} - y_k) + Ay_k = f,\\ (\omega_{k+1}^{(2)}E + A_2)(y_{k+1} - y_k) - (\omega_{k+1}^{(2)}E - A_1)(y_{k+1/2} - y_k) + Ay_k = f\end{gathered} \tag{5}$$

et de la première égalité ôtons la seconde. Il vient

$$y_{k+1/2} - y_k = (\omega_{k+1}^{(2)}E + A_2)\frac{y_{k+1} - y_k}{\omega_{k+1}^{(1)} + \omega_{k+1}^{(2)}} = (\omega_{k+1}^{(2)}E + A_2)\frac{y_{k+1} - y_k}{\tau_{k+1}}.$$

En portant cette expression dans (5), on obtient le schéma (2). Le passage inverse est évident.

Pour trouver y_{k+1} il est possible d'utiliser un autre algorithme, en traitant (2) comme un schéma avec correction ω_k,

$$(\omega_{k+1}^{(1)}E + A_1)\,v = r_k, \quad r_k = Ay_k - f,$$
$$(\omega_{k+1}^{(2)}E + A_2)\,w_k = v,$$
$$y_{k+1} = y_k - \tau_{k+1}w_k, \quad k = 0,\ 1,\ \ldots$$

Cet algorithme, comparé à (4), est plus économique, mais il exige toutefois une mémorisation plus importante d'information intermédiaire, c'est-à-dire qu'il a besoin d'une mémoire supplémentaire de l'ordinateur, ce qui n'est pas toujours commode.

Notons que lors de l'élaboration du schéma itératif (3), comme au cours de la construction des algorithmes, aucune condition n'était imposée aux opérateurs A_1 et A_2, hormis l'hypothèse naturelle de la non-dégénérescence des opérateurs $\omega_k^{(\alpha)} E + A_\alpha$, $\alpha = 1,\ 2$. Toutes les autres exigences envers les opérateurs A_1 et A_2 sont en rapport avec le problème du choix optimal des paramètres $\omega_k^{(1)}$ et $\omega_k^{(2)}$.

2. Position du problème du choix des paramètres. Dans la méthode des directions alternées on a affaire à deux suites de paramètres $\{\omega_k^{(1)}\}$ et $\{\omega_k^{(2)}\}$ qui seront choisies sur la base de la condition du minimum de la norme de l'opérateur résolvant dans l'espace de départ H.

Pour résoudre le problème du choix des paramètres d'itération, il faut énoncer certaines hypothèses sur les opérateurs A_1 et A_2, de caractère fonctionnel, et donner une certaine information à priori. Formulons ces hypothèses.

On admettra que l'opérateur A peut être représenté sous forme de somme de deux opérateurs A_1 et A_2 autoadjoints et permutables:

$$A = A_1 + A_2, \quad A_1 = A_1^*, \quad A_2 = A_2^*, \quad A_1A_2 = A_2A_1. \tag{6}$$

Supposons que l'information à priori est donnée sous forme de bornes δ_α et Δ_α de l'opérateur A_α, $\alpha = 1,\ 2$, etc.

$$\delta_1 E \leqslant A_1 \leqslant \Delta_1 E, \quad \delta_2 E \leqslant A_2 \leqslant \Delta_2 E, \tag{7}$$

avec la satisfaction de la condition

$$\delta_1 + \delta_2 > 0. \tag{8}$$

Notons que de (6)-(8) il s'ensuit que l'opérateur A est autoadjoint et défini positif.

Si l'hypothèse sur la permutabilité des opérateurs A_1 et A_2 est vérifiée, on dira qu'on a affaire à un cas *de commutativité* ou à un *cas général.* La condition (6) garantit que les opérateurs B_k sont autoadjoints pour tout k. De fait, en vertu de (6), les opérateurs $\omega_k^{(1)}E + A_1$ et $A_2 + \omega_k^{(2)}E$ sont autoadjoints et permutables, tandis que le produit des opérateurs autoadjoints et permutables est un opérateur autoadjoint.

Passons à l'étude de la convergence du schéma itératif (2). En portant $y_k = z_k + u$, où z_k est l'erreur et u la solution de l'équation (1), dans (2), on obtient pour z_k une équation homogène

$$z_{k+1} = S_{k+1} z_k, \quad k = 0, 1, \ldots, \quad z_0 = y_0 - u, \tag{9}$$

où

$$S_k = E - \tau_k B_k^{-1} A =$$
$$= (\omega_k^{(2)} E + A_2)^{-1} (\omega_k^{(1)} E + A_1)^{-1} (\omega_k^{(2)} E - A_1)(\omega_k^{(1)} E - A_2). \tag{10}$$

Utilisant (9), exprimons z_n au moyen z_0. Il vient

$$z_n = T_{n,0} z_0, \quad T_{n,0} = \prod_{j=1}^{n} S_j = S_n S_{n-1} \ldots S_1, \tag{11}$$

où $T_{n,0}$ est l'opérateur résolvant. Vu que les opérateurs A_1 et A_2 sont permutables, l'ordre des facteurs communs dans (10) est quelconque, tous les opérateurs S_k étant autoadjoints et permutables par paire, l'opérateur $T_{n,0}$ est donc autoadjoint dans H: $T_{n,0} = R_n(A_1, A_2)$, où $R_n(x, y)$ est un produit des fonctions de fractions rationnelles de x et y:

$$R_n(x, y) = \prod_{j=1}^{n} \frac{\omega_j^{(2)} - x}{\omega_j^{(1)} + x} \frac{\omega_j^{(1)} - y}{\omega_j^{(2)} + y}. \tag{12}$$

De (11) on obtient

$$\| z_n \| \leqslant \| T_{n,0} \| \, \| z_0 \|. \tag{13}$$

Comme l'opérateur $T_{n,0}$ est autoadjoint, on a $T_{n,0} = \max\limits_k |\lambda_k(T_{n,0})|$, où $\lambda_k(T_{n,0})$ sont les valeurs propres de l'opérateur $T_{n,0}$. Ensuite, en vertu des conditions (6) (voir point 5, § 1, ch. V), les opérateurs A_1, A_2 et $T_{n,0}$ possèdent un système commun de fonctions propres. Par suite,

$$\lambda_k(T_{n,0}) = R_n(\lambda_{k_1}^{(1)}, \lambda_{k_2}^{(2)}),$$

où $\lambda_{k_1}^{(1)}$ et $\lambda_{k_2}^{(2)}$ sont les valeurs propres des opérateurs A_1 et A_2 respectivement, de plus, en vertu de (7), on a $\delta_1 \leqslant \lambda_{k_1}^{(1)} \leqslant \Delta_1$, $\delta_2 \leqslant \lambda_{k_2}^{(2)} \leqslant \Delta_2$. Par conséquent,

$$\| T_{n,0} \| = \max_{k_1, k_2} |R_n(\lambda_{k_1}^{(1)}, \lambda_{k_2}^{(2)})| \leqslant \max_{\substack{\delta_1 \leqslant x \leqslant \Delta_1 \\ \delta_2 \leqslant y \leqslant \Delta_2}} |R_n(x, y)|.$$

En portant cette estimation dans (13), il vient

$$\| z_n \|_D \leqslant \max_{\substack{\delta_1 \leqslant x \leqslant \Delta_1 \\ \delta_2 \leqslant y \leqslant \Delta_2}} |R_n(x, y)| \, \| z_0 \|_D, \tag{14}$$

où $R_n(x, y)$ est défini dans (12), tandis que $D = E$. Notons qu'en vertu de la permutabilité des opérateurs A_1 et A_2 l'opérateur $T_{n,0}$ sera autoadjoint également dans l'espace énergétique H_D pour

$D = A, A^2$. Aussi, en vertu du lemme 5, § 1, ch. V, aura-t-on $\| T_{n,0} \| = \| T_{n,0} \|_A = \| T_{n,0} \|_A$, et, par suite, l'estimation (14) se vérifie pour $D = A$, $D = A^2$.

Bref, le problème de l'appréciation de l'erreur du schéma itératif (2) se réduit au problème d'appréciation du maximum du module de la fonction de deux variables $R_n(x, y)$ dans le rectangle $G = \{\delta_1 \leqslant x \leqslant \Delta_1, \delta_2 \leqslant y \leqslant \Delta_2\}$ et au choix des paramètres d'itération sur la base de la condition du minimum du maximum du module de cette fonction. Le problème ainsi posé est suffisamment compliqué; on le réduira au point 3 à un problème plus simple consistant dans la recherche de la fonction de la fraction rationnelle d'une variable s'écartant le moins de zéro sur le segment.

3. Transformation en fraction linéaire. Etudions la fonction $R(x, y)$. Avec la transformation en fraction linéaire des inconnues, appliquons le rectangle G sur un carré $\{\eta \leqslant u \leqslant 1, \eta \leqslant v \leqslant 1, \eta > 0\}$, la transformation étant choisie de la sorte qu'elle ne modifie pas la forme de la fonction $R_n(x, y)$. La transformation cherchée est de la forme

$$x = \frac{ru - s}{1 - tu}, \quad y = \frac{rv + s}{1 + tv}, \quad \eta \leqslant u, \quad v \leqslant 1, \tag{15}$$

où r, s, t et η doivent être définis.

En portant (15) dans (12) et en introduisant de nouveaux paramètres $\varkappa_j^{(1)}$ et $\varkappa_j^{(2)}$,

$$\varkappa_j^{(1)} = \frac{\omega_j^{(1)} - s}{r - t\omega_j^{(1)}}, \quad \varkappa_j^{(2)} = \frac{\omega_j^{(2)} + s}{r + t\omega_j^{(2)}}, \quad j = 1, 2, \ldots, n, \tag{16}$$

on obtient

$$R_n(x, y) = P_n(u, v) = \prod_{j=1}^{n} \frac{\varkappa_j^{(2)} - u}{\varkappa_j^{(1)} + u} \, \frac{\varkappa_j^{(1)} - v}{\varkappa_j^{(2)} + v}.$$

De (16) tirons les rapports à l'aide desquels les paramètres $\omega_j^{(1)}$ et $\omega_j^{(2)}$ s'expriment au moyen des paramètres introduits $\varkappa_j^{(1)}$ et $\varkappa_j^{(2)}$:

$$\omega_j^{(1)} = \frac{r\varkappa_j^{(1)} + s}{1 - t\varkappa_j^{(1)}}, \quad \omega_j^{(2)} = \frac{r\varkappa_j^{(2)} - s}{1 - t\varkappa_j^{(2)}}, \quad j = 1, 2, \ldots, n. \tag{17}$$

Bref, une fois les paramètres $\varkappa_j^{(1)}$ et $\varkappa_j^{(2)}$ trouvés, on est en mesure de déterminer les paramètres $\omega_j^{(1)}$ et $\omega_j^{(2)}$ à l'aide des formules (17).

La substitution (15) permet de passer au problème de la recherche des valeurs des paramètres $\varkappa_j^{(1)}$ et $\varkappa_j^{(2)}$ pour lesquelles on aboutit à

$$\min_{\varkappa^{(1)}, \varkappa^{(2)}} \max_{\eta \leqslant u, v \leqslant 1} |P_n(u, v)|.$$

Notons que si l'on impose certaines restrictions au choix des paramètres $\varkappa_j^{(1)}$ et $\varkappa_j^{(2)}$, par exemple, si l'on pose $\varkappa_j^{(1)} = \varkappa_j^{(2)} = \varkappa_j$, le minimum ne peut,

apparemment, que s'accroître. Donc

$$\min_{\varkappa^{(1)},\ \varkappa^{(2)}} \max_{\eta\leqslant u,\ v\leqslant 1} P_n(u, v)| \leqslant \min_{\varkappa} \max_{\eta\leqslant u,\ v\leqslant 1} \left| \prod_{j=1}^{n} \frac{\varkappa_j - u}{\varkappa_j + u} \frac{\varkappa_j - v}{\varkappa_j + v} \right| =$$

$$= \min_{\varkappa} \max_{\eta\leqslant u\leqslant 1} |r_n(u, \varkappa)|^2,\quad r_n(u, \varkappa) = \prod_{j=1}^{n} \frac{\varkappa_j - u}{\varkappa_j + u}.$$

Le problème du choix optimal des paramètres $\omega_j^{(1)}$ et $\omega_j^{(2)}$, posé plus haut, se réduit donc à la recherche de la fonction d'une fraction rationnelle $r_n(u, \varkappa)$ qui s'écarte le moins du zéro sur le segment $[\eta, 1]$. Autrement dit, il faut trouver des $\varkappa_j^*$ tels que

$$\max_{\eta\leqslant u\leqslant 1} |r_n(u, \varkappa^*)| = \min_{\varkappa} \max_{\eta\leqslant u< 1} |r_n(u, \varkappa)| = \rho.$$

Si ces paramètres sont obtenus, il s'ensuivra de (14) l'estimation de l'erreur z_n $\| z_n \|_D \leqslant \rho^2 \| z_0 \|_D$, et la précision ε sera atteinte une fois posé $\rho^2 = \varepsilon$.

Le choix cherché des paramètres d'itération sera donné au point 4, ici on cherchera les constantes r, s, t et η de la transformation (15).

Si $r \neq ts$, la transformation est monotone en u et v et, par suite, la transformation inverse $u = (x + s)/(r + tx)$, $v = (y - s)/(r - ty)$ sera monotone en x et y. Aussi, pour l'application du rectangle $\{\delta_1 \leqslant x \leqslant \Delta_1,\ \delta_2 \leqslant y \leqslant \Delta_2\}$ sur le carré $\{\eta \leqslant u,\ v \leqslant 1\}$, suffit-il que les extrémités du segment $[\delta_\alpha, \Delta_\alpha]$ deviennent les extrémités du segment $[\eta, 1]$. On obtient ainsi quatre rapports permettant de déterminer les constantes de la transformation (15):

$$\delta_1 = \frac{r\eta - s}{1 - t\eta},\quad \delta_2 = \frac{r\eta + s}{1 + t\eta},\quad \Delta_1 = \frac{r - s}{1 - t},\quad \Delta_2 = \frac{r + s}{1 + t}. \tag{18}$$

Cherchons la solution du système non linéaire (18). Notons d'abord qu'en vertu de l'hypothèse (8) les inégalités

$$\Delta_2 + \delta_1 \geqslant \delta_1 + \delta_2 > 0,\quad \Delta_1 + \delta_2 \geqslant \delta_1 + \delta_2 > 0 \tag{19}$$

se vérifient. Ensuite, de (18) on tire

$$\begin{aligned} \Delta_1 - \delta_1 &= \frac{(1-\eta)(r - st)}{(1-t)(1-t\eta)},\quad \Delta_2 - \delta_2 = \frac{(1-\eta)(r-st)}{(1+t)(1+t\eta)},\\ \Delta_2 + \delta_1 &= \frac{(1+\eta)(r-st)}{(1+t)(1-t\eta)},\quad \Delta_1 + \delta_2 = \frac{(1+\eta)(r-st)}{(1-t)(1+t\eta)}. \end{aligned} \tag{20}$$

De là on obtient

$$\left(\frac{1-\eta}{1+\eta}\right)^2 = \frac{(\Delta_1 - \delta_1)(\Delta_2 - \delta_2)}{(\Delta_1 + \delta_2)(\Delta_2 + \delta_1)} < 1,$$

et, puisqu'en vertu de (19) le dénominateur ne devient pas nul, on a

$$\eta = \frac{1-a}{1+a},\quad a = \sqrt{\frac{(\Delta_1 - \delta_1)(\Delta_2 - \delta_2)}{(\Delta_1 + \delta_2)(\Delta_2 + \delta_1)}},\quad \eta \in [0, 1]. \tag{21}$$

Cherchons maintenant t. De (20) il vient

$$\frac{\Delta_2 + \delta_1}{\Delta_1 - \delta_1} = \frac{1+\eta}{1-\eta} \frac{1-t}{1+t} = \frac{1}{a} \frac{1-t}{1+t}.$$

De là on obtient

$$t=\frac{1-b}{1+b}, \quad b=\frac{\Delta_2+\delta_1}{\Delta_1-\delta_1}\,a. \tag{22}$$

Des deux dernières équations du système (18) on déduit

$$r=\frac{1}{2}\left[\Delta_1(1-t)+\Delta_2(1+t)\right]=\frac{1+t}{2}\left[\Delta_2+\Delta_1 b\right]=\frac{\Delta_2+\Delta_1 b}{1+b}, \tag{23}$$

$$s=\frac{1}{2}\left[\Delta_2(1+t)-\Delta_1(1+t)\right]=\frac{1+t}{2}\left[\Delta_2-\Delta_1 b\right]=\frac{\Delta_2-\Delta_1 b}{1+b}. \tag{24}$$

Vu que

$$r-st=\frac{2b(\Delta_1+\Delta_2)}{(1+b)^2}>0, \quad |t|<1,$$

la transformation (15) est effectivement monotone. On montrera au point 4 que $\eta < \varkappa_j = \varkappa_j^{(1)} = \varkappa_j^{(2)} < 1$. Donc dans (17) les dénominateurs ne s'annulent pas.

Voyons quelques exemples. Soient $\delta_1 = \delta_2 = \delta$ et $\Delta_1 = \Delta_2 = \Delta$, autrement dit les bornes des opérateurs A_1 et A_2 sont les mêmes. On a alors $\eta = \delta/\Delta$, $t = s = 0$, $r = \Delta$, $\omega_j^{(1)} = \omega_j^{(2)} = \Delta\varkappa_j$. Soient maintenant $\delta_1 = 0$, $\delta_2 = \delta$, $\Delta_1 = \Delta_2 = \Delta$, c'est-à-dire que l'opérateur A_1 est dégénéré. Alors

$$\eta=\delta/(\Delta+\sqrt{\Delta^2-\delta^2}), \quad t=\eta, \quad s=\Delta\eta, \quad r=\Delta,$$

$$\omega_j^{(1)}=\frac{\Delta\varkappa_j+\Delta\eta}{1+\eta\varkappa_j}, \quad \omega_j^{(2)}=\frac{\Delta\varkappa_j-\Delta\eta}{1-\eta\varkappa_j}, \quad j=1,\ 2,\ \ldots,\ n.$$

4. Choix optimal des paramètres. Donnons la solution du problème du choix optimal des paramètres d'itération. A la différence du cas de recherche du polynôme s'écartant le moins de zéro, qui a été étudié au § 2 du ch. VI, ici les paramètres d'itération $\varkappa_j$ s'expriment non pas sous forme de fonctions trigonométriques mais à l'aide des fonctions elliptiques de Jacobi.

Rappelons quelques définitions. L'intégrale définie

$$K(k)=\int_0^{\pi/2}\frac{d\varphi}{1-k^2\sin^2\varphi}$$

est dite *intégrale elliptique complète de première espèce*, le nombre k étant le *module* de cette intégrale et le nombre $k' = \sqrt{1-k^2}$ le *module complémentaire*. On a adopté les notations $K(k') = K'(k)$.

Si l'on désigne au moyen de $u(z, k)$ la fonction

$$u(z,\ k)=\int_z^1\frac{dy}{\sqrt{(1-y^2)(y^2-k^2)}},$$

alors la fonction $z = \mathrm{dn}\,(u, k')$, inverse de $u(z, k)$, est appelée *fonction elliptique de Jacobi* de l'argument u et du module k'.

En se servant de ces désignations, il est possible d'écrire la solution précise du choix optimal des paramètres d'itération $\varkappa_j$ de l'intégration sous la forme:

$$\varkappa_j \in \mathfrak{M}_n = \left\{ \mu_i = \mathrm{dn}\left(\frac{2i-1}{2n} K'(\eta), \eta' \right), \quad i = 1, 2, \ldots, n \right\}, \qquad (25)$$
$$j = 1, 2, \ldots, n,$$

où le nombre d'itérations n, suffisamment grand pour aboutir à la précision ε, est apprécié suivant la formule

$$n \geqslant n_0(\varepsilon) = \frac{1}{4} \frac{K'(\eta) K'(\varepsilon)}{K(\eta) K(\varepsilon)}. \qquad (26)$$

Ici, comme au cas de la méthode de Tchébychev, on choisit successivement en guise de $\varkappa_j$ tous les éléments de l'ensemble $\mathfrak{M}_n$. Formulons les résultats obtenus pour la méthode des directions alternées au cas de commutativité sous forme d'un théorème.

Théorème 1. *Supposons que les conditions* (6)-(8) *sont remplies et les paramètres* $\omega_j^{(1)}$ *et* $\omega_j^{(2)}$ *choisis suivant les formules*

$$\omega_j^{(1)} = \frac{r\varkappa_j + s}{2 + t\varkappa_j}, \quad \omega_j^{(2)} = \frac{r\varkappa_j - s}{1 - t\varkappa_j}, \quad j = 1, 2, \ldots, n, \qquad (27)$$

où $\varkappa_j$ *et* n *sont définis dans* (25), (26), *tandis que* r, s, t *et* η *dans* (21)-(24). *La méthode des directions alternées* (2), (3) *converge dans* H_D *et, après n itérations, on a pour l'erreur* $z_n = y_n - u$ *l'estimation* $\| z_n \|_D \leqslant \varepsilon \| z_0 \|_D$, *où* $D = E$, A *ou* A^2, *quant à* n, *il est défini selon* (26).

Passons à présent au problème des calculs impliqués par la mise en œuvre de la méthode des directions alternées avec un choix optimal des paramètres. Cherchons les formules approchées de calcul de $\varkappa_j$ et n et indiquons l'ordre à suivre dans le choix des paramètres $\varkappa_j$ à partir de l'ensemble $\mathfrak{M}_n$.

En profitant de la représentation asymptotique des intégrales elliptiques complètes pour des petites valeurs de k:

$$\frac{1}{K(k)} = \frac{2}{\pi} + O(k^2), \quad K'(k) = \ln\frac{4}{k} + O\left(k^2 \ln\frac{1}{k}\right),$$

on obtient de (26) la formule approchée suivante pour le nombre d'itérations n;

$$n \geqslant n_0(\varepsilon) = \frac{1}{\pi^2} \ln\frac{4}{\eta} \ln\frac{4}{\varepsilon}. \qquad (28)$$

Examinons maintenant la question du calcul de μ_i. La fonction dn (u, k') décroît de façon monotone en u en prenant les valeurs suivantes: dn $(0, k') = 1$, dn $(K'(k), k') = K$. Donc $\eta < \mu_n < \mu_{n-1} < \ldots < \mu_1 < 1$. Ensuite, de la propriété de la fonction elliptique dn (u, k'):

$$dn(u, k') = \frac{k}{dn(K'(k) - u, k')}$$

il s'ensuit l'égalité

$$\mu_i = \eta / \mu_{n+1-i}, \quad i = 1, 2, \ldots \qquad (29)$$

Il suffit donc de trouver la moitié des valeurs de μ_i, les valeurs restantes s'obtenant à partir de la relation (29).

La formule approchée de μ_i sera obtenue en recourant au développement de la fonction dn (u, k') en puissance de k. A cette fin exprimons la fonction dn $(\sigma K'(\eta), \eta')$ au moyen des fonctions thêta de Jacobi écrites en série. Il vient

$$\mathrm{dn}(\sigma K'(\eta), \eta') = \frac{\sqrt{\eta}\,\theta_3\left(\dfrac{i\sigma\pi K'}{K}, \bar{q}\right)}{\theta_2\left(\dfrac{i\sigma\pi K'}{K'}, \bar{q}\right)} = \sqrt{\eta}\,\bar{q}^{\frac{2\sigma-1}{4}} \frac{\sum\limits_{m=-\infty}^{\infty} \bar{q}^{m(m+\sigma)}}{\sum\limits_{m=-\infty}^{\infty} \bar{q}^{m(m-1+\sigma)}},$$

où $\bar{q} = \exp\left(-\dfrac{\pi K'(\eta)}{K(\eta)}\right) = \dfrac{\eta^2}{16}\left(1+\dfrac{\eta^2}{2}\right) + O(\eta^6)$.

De là on tire

$$\mathrm{dn}(\sigma K'(\eta), \eta') = \sqrt{\eta}\, q^{\frac{2\sigma-1}{4}} \frac{1+q^{1-\sigma}+q^{1+\sigma}}{1+q^{\sigma}+q^{2-\sigma}} + O(\eta^{\nu}), \tag{30}$$

où

$$q = \eta^2(1+\eta^2/2)/16, \quad \nu = \begin{cases} 4+5\sigma, & 0 < \sigma < 1/2, \\ 8-3\sigma, & 1/2 \leqslant \sigma < 1. \end{cases}$$

Pour $\sigma \geqslant 1/2$, l'ordre du terme résiduel dans (30) vaut 5 de façon uniforme en σ, tandis que pour $\sigma < 1/2$ l'ordre vaut 4. Donc la formule approchée pour dn $(\sigma K'(\eta), \eta')$ sera plus précise pour $\sigma \geqslant 1/2$ devant $\sigma < 1/2$.

De (25), (29) et (30), on obtient les formules suivantes permettant de calculer μ_i:

$$\mu_i = \sqrt{\eta}\, q^{\frac{2\sigma_i-1}{4}} \frac{1+q^{1-\sigma_i}+q^{1+\sigma_i}}{1+q^{\sigma_i}+q^{2-\sigma_i}}, \quad [n/2]+1 \leqslant i \leqslant n,$$

$$\mu_i = \eta/\mu_{n+1-i}, \quad 1 \leqslant i \leqslant [n/2], \quad \sigma_i = (2i-1)/(2n),$$

$$q = \eta^2(1+\eta^2/2)/16,$$

où $[a]$ est la partie entière de a.

Examinons maintenant la question de l'ordre du choix de $\varkappa_j$ à partir de l'ensemble $\mathfrak{M}_n$. De la définition de l'opérateur de passage S_j dans le schéma (2) et des propriétés (6), (7) il s'ensuit que

$$\| S_j \| = \max_k | \lambda_k(S_j) | \leqslant \max_{\substack{\delta_1 \leqslant x \leqslant \Delta_1 \\ \delta_2 \leqslant y \leqslant \Delta_2}} \left| \frac{\omega_j^{(2)}-x}{\omega_j^{(1)}+x} \, \frac{\omega_j^{(1)}-y}{\omega_j^{(2)}+y} \right|$$

ou, en vertu de la substitution (15),

$$\| S_j \| \leqslant \max_{\eta \leqslant u \leqslant 1} \left| \frac{\varkappa_j - u}{\varkappa_j + u} \right|^2.$$

Vu que tous les $\varkappa_j$ appartiennent à l'intervalle $(\eta, 1)$, il s'ensuit que $\| S_j \| < 1$ pour tout j. La méthode itérative (2), (3) sera donc stable vis-à-vis des erreurs d'arrondi pour tout ordre de choix de $\varkappa_j$ à partir de l'ensemble $\mathfrak{M}_n$, par exemple, $\varkappa_j = \mu_j$, $j = 1, 2, \ldots, n$.

En conclusion de ce point, montrons que pour le jeu de paramètres $\omega_j^{(1)}$ et $\omega_j^{(2)}$ construit les opérateurs $\omega_j^{(\alpha)} E + A_\alpha$, $\alpha = 1, 2$, sont définis positifs dans H pour tout j. En effet, à partir de (27) on obtient

$$\frac{\partial \omega_j^{(1)}}{\partial \varkappa_j} = \frac{r-st}{(1+t\varkappa_j)^2} > 0, \quad \frac{\partial \omega_j^{2}}{\partial \varkappa_j} = \frac{r-st}{(1-t\varkappa_j)^2} > 0.$$

Vu que les dénominateurs dans (27) ne s'annulent pas et que $\eta < \varkappa_j < 1$, il en suit, ainsi que de (18), que

$$\delta_2 = \frac{r\eta+s}{1+t\eta} \leqslant \omega_j^{(1)} \leqslant \frac{r+s}{1+t} = \Delta_2, \quad \delta_1 = \frac{r\eta-s}{1-t\eta} \leqslant \omega_j^{(2)} \leqslant \frac{r-s}{1-t} = \Delta_1. \quad (31)$$

Par conséquent, en vertu de l'hypothèse (7), on obtient de (31)

$$\omega_j^{(1)} E + A_1 \geqslant (\delta_1 + \delta_2) E, \quad \omega_j^{(2)} E + A_2 \geqslant (\delta_1 + \delta_2) E,$$

et comme selon l'hypothèse (8) $\delta_1 + \delta_2 > 0$, la proposition est démontrée.

§ 2. Exemples d'application de la méthode

1. Problème discret de Dirichlet pour l'équation de Poisson dans un rectangle. On commencera l'étude des exemples de mise en œuvre de la méthode des directions alternées avec la recherche de la solution du problème discret de Dirichlet pour l'équation de Poisson dans un rectangle.

Soit qu'il s'agit de rechercher sur le maillage rectangulaire $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, \ 0 \leqslant i \leqslant N_1, \ 0 \leqslant j \leqslant N_2, \ h_\alpha = l_\alpha / N_\alpha, \ \alpha = 1, 2\}$, introduit dans un rectangle $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha, \ \alpha = 1, 2\}$, la solution du problème

$$\Lambda y = (\Lambda_1 + \Lambda_2) y = -\varphi(x), \quad x \in \omega, \quad y(x) = g(x), \quad x \in \gamma, \qquad \Lambda_\alpha y = y_{\bar{x}_\alpha x_\alpha}, \quad \alpha = 1, 2. \quad (1)$$

Désignons par H l'espace des fonctions de mailles associées à ω avec le produit scalaire

$$(u, v) = \sum_{x \in \omega} u(x) v(x) h_1 h_2.$$

Définissons sur H les opérateurs A, A_1 et A_2 de la façon suivante: $Ay = -\Lambda \mathring{y}$, $A_\alpha y = -\Lambda_\alpha \mathring{y}$, $\alpha = 1, 2$, où $y \in H$, $\mathring{y} \in \mathring{H}$, $y(x) = \mathring{y}(x)$ pour $x \in \omega$, $\mathring{H}$ étant l'ensemble des fonctions de mailles associées à $\bar{\omega}$ et s'annulant sur γ.

Le problème de différences (1) peut alors être écrit sous forme d'une équation opératorielle $Au = f$, où $A = A_1 + A_2$.

Comme on le sait (voir point 5, § 1, ch. V), les opérateurs A_α sont autoadjoints dans H et ont des bornes δ_α et Δ_α:

$$\delta_\alpha = \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2 l_\alpha}, \quad \Delta_\alpha = \frac{4}{h_\alpha^2} \cos^2 \frac{\pi h_\alpha}{2 l_\alpha}, \quad \alpha = 1, 2,$$

qui coïncident avec les valeurs propres, minimale et maximale, des opérateurs de différences Λ_α. Il reste à vérifier la permutabilité

des opérateurs A_1 et A_2. En utilisant la définition des opérateurs A_α et des opérateurs de différences Λ_α, on obtient

$$A_1A_2y = \overset{\circ}{y}_{\bar{x}_1x_1\bar{x}_2x_2} = \overset{\circ}{y}_{\bar{x}_2x_2\bar{x}_1x_1} = A_2A_1y,$$

ce qu'il fallait démontrer.

Bref, les conditions exigées pour la mise en œuvre de la méthode des directions alternées au cas de commutativité sont satisfaites pour l'exemple considéré.

En se servant de la définition des opérateurs A_1 et A_2, on peut écrire l'algorithme de la méthode des directions alternées pour l'exemple considéré de la façon suivante:

$$\begin{gathered} \omega_{k+1}^{(1)}y_{k+1/2} - \Lambda_1 y_{k+1/2} = \omega_{k+1}^{(1)}y_k + \Lambda_2 y_k + \varphi, \quad h_1 \leqslant x_1 \leqslant l_1 - h_1, \\ y_{k+1/2}(x) = g(x), \quad x_1 = 0, \ l_1, \quad h_2 \leqslant x_2 \leqslant l_2 - h_2, \end{gathered} \tag{2}$$

$$\begin{gathered} \omega_{k+1}^{(2)}y_{k+1} - \Lambda_2 y_{k+1} = \omega_{k+1}^{(2)}y_{k+1/2} + \Lambda_1 y_{k+1/2} + \varphi, \quad h_2 \leqslant x_2 \leqslant l_2 - h_2, \\ y_{k+1}(x) = g(x), \quad x_2 = 0, \ l_2, \quad h_1 \leqslant x_1 \leqslant l_1 - h_1, \end{gathered} \tag{3}$$

en outre, $y_k(x) = g(x)$ avec $x \in \gamma$ pour tout $k \geqslant 0$. Donc pour déterminer $y_{k+1/2}$ sur ω, l'algorithme de la méthode consiste dans la résolution successive pour chaque x_2 fixé des problèmes aux limites triponctuels (2) suivant la direction de x_1, et pour déterminer la nouvelle approximation itérative y_{k+1} sur ω, dans la résolution pour chaque x_1 des problèmes aux limites (3) suivant la direction de x_2. L'alternation des directions suivant lesquelles sont résolus les problèmes aux limites (2), (3) fut à l'origine de l'appellation de la méthode (méthode des directions alternées).

Pour la résolution des problèmes (2), (3), on peut recourir à la méthode du balayage. Ecrivons les équations (2), (3) sous forme de système triponctuel et vérifions si les conditions suffisantes de stabilité de la méthode du balayage sont satisfaites. Les équations prendront la forme

$$\begin{gathered} -y_{k+1/2}(i+1, j) + (2 + h_1^2\omega_{k+1}^{(1)})\, y_{k+1/2}(i, j) - \\ - y_{k+1/2}(i-1, j) = \varphi_1(i, j), \quad 1 \leqslant i \leqslant N_1 - 1, \\ y_{k+1/2}(0, j) = g(0, j), \quad y_{k+1/2}(N_1, j) = g(N_1, j), \\ 1 \leqslant j \leqslant N_2 - 1, \end{gathered} \tag{4}$$

où

$$\varphi_1(i, j) = \frac{h_1^2}{h_2^2}\,[y_k(i, j+1) - (2 + h_2^2\omega_{k+1}^{(1)})\, y_k(i, j) + y_k(i, j-1) + h_2^2\varphi(i, j)];$$

$$\begin{gathered} -y_{k+1}(i, j+1) + (2 + h_2^2\omega_{k+1}^{(2)})\, y_{k+1}(i, j) - y_{k+1}(i, j-1) = \varphi_2(i, j), \\ 1 \leqslant j \leqslant N_2 - 1, \\ y_{k+1}(i, 0) = g(i, 0), \quad y_{k+1}(i, N_2) = g(i, N_2), \quad 1 \leqslant i \leqslant N_1 - 1, \end{gathered} \tag{5}$$

où

$$\varphi_2(i,\ j)=\frac{h_2^2}{h_1^2}\,[y_{k+1/2}(i+1,\ j)-(2+h_1^2\omega_{k+1}^{(2)})\,y_{k+1/2}(i,\ j)+ \\ +y_{k+1/2}(i-1,\ j)+h_1^2\varphi(i,\ j)].$$

Vu que pour l'exemple donné $\delta_1>0$, $\delta_2>0$, les paramètres $\omega_k^{(1)}$ et $\omega_k^{(2)}$ sont positifs en vertu de l'inégalité (31) du point 4, § 1. Donc dans les équations triponctuelles (4) et (5) les coefficients associés à $y_{k+1/2}(i,\ j)$ et $y_{k+1}(i,\ j)$ dominent sur les autres coefficients. Par conséquent, la méthode du balayage, appliquée aux problèmes (4), (5), sera stable vis-à-vis des erreurs d'arrondi.

Calculons le nombre d'opérations arithmétiques qu'il faut effectuer pour la mise en œuvre d'une seule itération dans la méthode (2), (3) appliquée à l'exemple concerné. Il suffit de calculer le nombre d'opérations pour le problème (4), pour le problème (5) le calcul étant mené de façon analogue.

Les formules de la méthode du balayage prennent pour le problème (4) la forme suivante (j est fixé)

$$y_{k+1/2}(i,\ j)=\alpha_i y_{k+1/2}(i+1,\ j)+\beta_i,\quad 1\leqslant i\leqslant N_1-1,$$
$$y_{k+1/2}(N_1,\ j)=g(N_1,\ j),$$
$$\alpha_{i+1}=1/(C-\alpha_i),\ i=1,\ 2,\ \ldots,\ N_1-1,\quad \alpha_1=0,\ C=2+h_1^2\omega_{k+1}^{(1)},$$
$$\beta_{i+1}=\alpha_{i+1}(\varphi_1(i,\ j)+\beta_i),$$
$$i=1,\ 2,\ \ldots,\ N_1-1,\ \beta_1=g(0,\ j).$$

Notons que les coefficients de balayage α_i ne dépendent pas de j et peuvent donc être calculés une seule fois avec $2(N_1-1)$ opérations arithmétiques. Ensuite, pour le calcul de $\varphi_1(i,\ j)$ sur le maillage ω, il faut $6(N_1-1)(N_2-1)$ opérations arithmétiques. Les coefficients de balayage β_i et la solution $y_{k+1/2}$ doivent être recalculés à chaque j. Il faut pour cela $4(N_1-1)(N_2-1)$ opérations. En tout, l'obtention de $y_{k+1/2}$ sur le maillage ω pour y_k donné se soldera par $Q_1=10(N_1-1)(N_2-1)+2(N_1-1)$ opérations arithmétiques. Pour obtenir y_{k+1} de (15), une fois $y_{k+1/2}$ calculé, il faut dépenser $Q_2=10(N_1-1)(N_2-1)+2(N_2-1)$ opérations. Bref, pour l'exemple considéré, la mise en œuvre d'une seule itération s'effectue en

$$Q=20(N_1-1)(N_2-1)+2(N_1-1)+2(N_2-1)\qquad(6)$$

opérations arithmétiques.

Apprécions maintenant le nombre d'itérations n suffisant pour l'obtention de la solution avec la précision ε imposée. Dans le cas particulier, quand le domaine $\overline{G}$ est un carré de côté l ($l_1=l_2=l$)

et le maillage $\bar{\omega}$ est carré avec $N_1 = N_2 = N$ $(h_1 = h_2 = l/N)$, on a

$$\delta_1 = \delta_2 = \delta = \frac{4}{h^2} \sin^2 \frac{\pi h}{2l}, \quad \Delta_1 = \Delta_2 = \Delta = \frac{4}{h^2} \cos^2 \frac{\pi h}{2l}.$$

A partir de (21) et (28), § 1, on tire l'estimation suivante pour le nombre d'itérations:

$$n \geqslant n_0(\varepsilon) = 0{,}1 \ln \frac{4}{\eta} \ln \frac{4}{\varepsilon}, \quad \eta = \delta/\Delta = \mathrm{tg}^2 \frac{\pi h}{2l}$$

ou pour des h petits

$$n_0(\varepsilon) = 0{,}2 \ln (4N/\pi) \ln (4/\varepsilon), \tag{7}$$

c'est-à-dire que le nombre d'itérations est proportionnel au logarithme du nombre d'inconnues N suivant une direction.

De (6), (7) on obtient l'estimation suivante pour le nombre d'opérations arithmétiques $Q(\varepsilon)$ dépensées à la recherche de la solution du problème de différences (1) par la méthode des directions alternées avec la précision ε:

$$Q(\varepsilon) = nQ = 4N^2 \ln (4N/\pi) \ln (4/\varepsilon). \tag{8}$$

Pour comparer cette méthode avec la méthode directe de réduction totale (voir § 3, ch. III), adoptons dans (8) au lieu des logarithmes naturels les logarithmes de base 2.

On obtient

$$Q(\varepsilon) \approx 2{,}12\ N^2 \log_2 (4N/\pi) \log_2 (4/\varepsilon).$$

Vu que l'erreur d'approximation du schéma aux différences (1) est $O(h^2)$, il est logique de choisir ε égal à $O(h^2)$.

Si l'on pose $\varepsilon = 4/N^2$, il vient

$$Q(\varepsilon) = 4{,}24 N^2 \log_2 N \log_2 (4N/\pi).$$

Pour $N = 64$ on obtient $\varepsilon \approx 10^{-3}$ et

$$Q(\varepsilon) \approx 27{,}6\ N^2 \log_2 N.$$

La confrontation avec l'estimation du nombre d'opérations obtenu pour la méthode de réduction totale montre que pour le maillage indiqué la méthode des directions alternées exige environ 5,5 fois plus d'opérations arithmétiques que la méthode de réduction totale. Avec l'accroissement de N et la diminution de ε cette différence s'accentue.

Pour le cas particulier considéré, donnons le nombre d'itérations n en fonction du nombre de nœuds N suivant une direction pour $\varepsilon = 10^{-4}$.

A titre de comparaison, fournissons le nombre d'itérations obtenu pour les autres méthodes étudiées plus haut.

Tableau 11

N	Méthode itérative simple	Méthode explicite de Tchébychev	Méthode de surrelaxation	Méthode triangulaire alternée	Méthode des directions alternées
32	1909	101	65	16	8
64	7642	202	128	23	10
128	30577	404	257	32	11

Il s'ensuit du tableau que le plus petit nombre d'itérations est exigé par la méthode des directions alternées. Ce nombre d'itérations est supérieur à celui de la méthode triangulaire alternée avec paramètres de Tchébychev construite et étudiée au ch. X.

R e m a r q u e. Si au problème (1) étudié on applique la méthode des directions alternées avec paramètres constants, c'est-à-dire si $\omega_j^{(1)} \equiv \omega^{(1)}$, $\omega_j^{(2)} \equiv \omega^{(2)}$, $\tau_j = \omega^{(1)} + \omega^{(2)}$, on obtiendra alors de la formule (25), § 1, en vertu de l'égalité dn $\left(\frac{1}{2}K'(k),\ k'\right) = \sqrt{k}$, que $\varkappa_j \equiv \sqrt{\eta}$. Dans le cas particulier, quand $\delta_1 = \delta_2 = \delta$, $\Delta_1 = \Delta_2 = \Delta$, on a obtenu auparavant dans le point 3, § 1, la relation suivante liant les paramètres $\omega_j^{(1)}$, $\omega_j^{(2)}$ et $\varkappa_j$: $\omega_j^{(1)} = \omega_j^{(2)} = \Delta\varkappa_j$. Vu que dans ce cas $\eta = \delta/\Delta$, on en déduit que $\omega^{(1)} = \omega^{(2)} = \sqrt{\delta\Delta}$.

2. Troisième problème aux limites pour l'équation elliptique à variables séparables. Supposons que dans le rectangle $\overline{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha,\ \alpha = 1, 2\}$ il s'agit de trouver la solution du problème aux limites suivant:

$$Lu = \sum_{\alpha=1}^{2} \frac{\partial}{\partial x_\alpha}\left(k_\alpha(x_\alpha)\frac{\partial u}{\partial x_\alpha}\right) - qu = -f(x),\quad x \in G,$$

$$k_\alpha(x_\alpha)\frac{\partial u}{\partial x_\alpha} = \varkappa_{-\alpha}u - g_{-\alpha}(x),\quad x_\alpha = 0, \tag{9}$$

$$-k_\alpha(x_\alpha)\frac{\partial u}{\partial x_\alpha} = \varkappa_{+\alpha}u - g_{+\alpha}(x),\quad x_\alpha = l_\alpha,\quad \alpha = 1, 2.$$

Admettons que les conditions suivantes sont remplies:

$$0 \leqslant c_{1,\alpha} \leqslant k_\alpha(x_\alpha) \leqslant c_{2,\alpha},\quad \varkappa_{\pm\alpha} = \text{const} \geqslant 0,\quad \alpha = 1, 2, \tag{10}$$

$$q = \text{const} \geqslant 0,\quad \sum_{\alpha=1}^{2} \varkappa_{\pm\alpha}^2 + q^2 \neq 0$$

Le problème aux limites de Neumann ($\varkappa_{\pm\alpha} = 0$) pour le cas de $q = 0$ sera étudié séparément au ch. XII. Les conditions (10) garantissent l'existence et l'unicité de la solution du problème (9).

Sur un maillage rectangulaire $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G},\ 0 \leqslant i \leqslant N_1,\ 0 \leqslant j \leqslant N_2,\ h_\alpha = l_\alpha/N_\alpha,\ \alpha = 1, 2\}$ au problème (9) correspond le problème de différences aux limites

$$\Lambda y = (\Lambda_1 + \Lambda_2)\, y = -\varphi(x), \quad x \in \bar{\omega}, \tag{11}$$

où les opérateurs de différences Λ_1 et Λ_2 et le second membre φ se définissent de la façon suivante:

$$\Lambda_\alpha y = \begin{cases} \dfrac{2}{h_\alpha} a_\alpha(h_\alpha)\, y_{x_\alpha} - \left(0{,}5q + \dfrac{2}{h_\alpha}\varkappa_{-\alpha}\right) y, & x_\alpha = 0, \\ (a_\alpha(x_\alpha)\, y_{\bar{x}_\alpha})_{x_\alpha} - 0{,}5qy, & h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ -\dfrac{2}{h_\alpha} a_\alpha(l_\alpha)\, y_{\bar{x}_\alpha} - \left(0{,}5q + \dfrac{2}{h_\alpha}\varkappa_{+\alpha}\right) y, & x_\alpha = l_\alpha \end{cases}$$

pour $0 \leqslant x_\beta \leqslant l_\beta$, $\beta = 3 - \alpha$, $\alpha = 1, 2$ et $\varphi = f + \varphi_1 + \varphi_2$,

$$\varphi_\alpha(x) = \begin{cases} \dfrac{2}{h_\alpha} g_{-\alpha}(x), & x_\alpha = 0, \\ 0, & h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ \dfrac{2}{h_\alpha} g_{+\alpha}(x), & x_\alpha = l_\alpha. \end{cases}$$

Désignons par H l'espace des fonctions de mailles associées à $\bar{\omega}$, dont le produit scalaire se définit par la formule

$$(u, v) = \sum_{x \in \bar{\omega}} u(x)\, v(x)\, \hbar_1(x_1)\, \hbar_2(x_2),$$

$$\hbar_\alpha(x_\alpha) = \begin{cases} 0{,}5h_\alpha, & h_\alpha = 0, \\ h_\alpha, & h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha. \end{cases}$$

Les opérateurs A, A_1 et A_2 seront déterminés sur H par les relations $Ay = -\Lambda y$, $A_\alpha y = -\Lambda_\alpha y$, $\alpha = 1, 2$. On a montré au § 2, ch. V, que les opérateurs A_1 et A_2 ainsi définis sont autoadjoints et permutables. En outre, en vertu des conditions (10), l'opérateur A est défini positif dans H (c'est-à-dire $\delta_1 + \delta_2 > 0$). Il reste à trouver les bornes des opérateurs A_1 et A_2, autrement dit les constantes δ_α et Δ_α des inégalités $\delta_\alpha E \leqslant A_\alpha \leqslant \Delta_\alpha E$, $\alpha = 1, 2$.

Cherchons d'abord δ_α. De la définition des opérateurs A_α et des formules aux différences de Green, il vient

$$(A_\alpha y, y) = -\sum_{x_\beta=0}^{l_\beta}\sum_{x_\alpha=h_\alpha}^{l_\alpha-h_\alpha}[(a_\alpha y_{\bar{x}_\alpha})_{x_\alpha} - 0{,}5qy]\,y h_1 \hbar_2 -$$

$$-\sum_{x_\beta=0}^{l_\beta}\left[a_\alpha(h_\alpha)\,y_{x_\alpha} - \left(\varkappa_{-\alpha} + \frac{h_1}{4}q\right)y\right]y\Big|_{x_\alpha=0}\hbar_2 +$$

$$+\sum_{x_\beta=0}^{l_\beta}\left[a_\alpha(l_\alpha)\,y_{\bar{x}_\alpha} + \left(\varkappa_{+\alpha} + \frac{h_1}{4}q\right)y\right]y\Big|_{x_\alpha=l_\alpha}\hbar_2 =$$

$$= \sum_{x_\beta=0}^{l_\beta}\sum_{x_\alpha=h_\alpha}^{l_\alpha} a_\alpha y^2_{\bar{x}_\alpha} h_1 \hbar_2 + \sum_{x_\beta=0}^{l_\beta}\left(\varkappa_{-\alpha}y^2|_{x_\alpha=0} + \varkappa_{+\alpha}y^2|_{x_\alpha=l_\alpha}\right)\hbar^2 +$$

$$+0{,}5q\,(y^2, 1).$$

De là on tire que si $q = \varkappa_{-\alpha} = \varkappa_{+\alpha} = 0$, alors $\delta_\alpha = 0$. Si au moins une des grandeurs q, $\varkappa_{-\alpha}$ ou $\varkappa_{+\alpha}$ est différente de zéro, on peut obtenir δ_α de la façon suivante. En vertu du lemme 16, § 2, ch. V, on a

$$(y^2, 1)_{\bar{\omega}_\alpha} \leqslant \max_{x_\alpha \in \bar{\omega}_\alpha} v^\alpha(x_\alpha)\,(A_\alpha y, y)_{\bar{\omega}_\alpha}, \tag{12}$$

où $v^\alpha(x_\alpha)$ est la solution du problème aux limites triponctuel

$$\begin{gathered}(a_\alpha(x_\alpha)\,v^\alpha_{\bar{x}_\alpha})_{x_\alpha} - 0{,}5qv = -1, \quad h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ \frac{2}{h_\alpha}a_\alpha(h_\alpha)\,v^\alpha_{x_\alpha} - \left(0{,}5q + \frac{2}{h_\alpha}\varkappa_{-\alpha}\right)v^\alpha = -1, \; x_\alpha = 0, \\ -\frac{2}{h_\alpha}(a_\alpha(l_\alpha)\,v^\alpha_{\bar{x}_\alpha} - \left(0{,}5q + \frac{2}{h_\alpha}\varkappa_{+\alpha}\right)v^\alpha = -1, \; x_\alpha = l_\alpha,\end{gathered} \tag{13}$$

quant au produit scalaire, on le détermine de la façon suivante:

$$(u, v)_{\bar{\omega}_\alpha} = \sum_{x_\alpha=0}^{l_\alpha} u(x)\,v(x)\,\hbar_\alpha(x_\alpha).$$

En multipliant (12) par $\hbar_\beta(x_\beta)$ et en sommant en x_β de 0 à l_β, il vient

$$(y^2, 1) \leqslant \max_{x_\alpha \in \bar{\omega}_\alpha} v^\alpha(x_\alpha)\,(A_\alpha y, y)$$

et, par conséquent,

$$\delta_\alpha = \frac{1}{\max\limits_{x_\alpha \in \bar{\omega}_\alpha} v^\alpha(x_\alpha)}, \quad \alpha = 1, 2.$$

Cherchons maintenant Δ_α. A l'opérateur A_α correspond une matrice tridiagonale a_α. Désignons par $\mathscr{D}$ la partie diagonale de la matrice A_α, c'est-à-dire posons $\mathscr{D}y = d_\alpha(x_\alpha)\, y$,

$$d_\alpha(x_\alpha) = \begin{cases} 0{,}5q + \dfrac{2}{h_\alpha}\varkappa_{-\alpha} + \dfrac{2}{h_\alpha^2} a_\alpha(h_\alpha), & x_\alpha = 0, \\ 0{,}5q + \dfrac{1}{h_\alpha^2}(a_\alpha(x_\alpha) + a_\alpha(x_\alpha + h_\alpha)), & h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ 0{,}5q + \dfrac{2}{h_\alpha}\varkappa_{+\alpha} + \dfrac{2}{h_\alpha^2} a_\alpha(l_\alpha), & x_\alpha = l_\alpha. \end{cases}$$

Étudions le problème aux valeurs propres

$$A_\alpha y - \lambda \mathscr{D} y = 0, \quad x \in \bar{\omega}. \tag{14}$$

On montre sans peine que si λ est une valeur propre du problème (14), $2 - \lambda$ est également une valeur propre. Par conséquent,

$$\lambda_{\min}\mathscr{D} \leqslant A_\alpha \leqslant (2 - \lambda_{\min})\,\mathscr{D}$$

ou

$$(A_\alpha y, y) \leqslant (2 - \lambda_{\min})(\mathscr{D}y, y) \leqslant (2 - \lambda_{\min}) \max_{x_\alpha \in \bar{\omega}_\alpha} d_\alpha(x_\alpha)(y, y).$$

Aussi en guise de Δ_α peut-on prendre

$$\Delta_\alpha = (2 - \lambda_{\min}) \max_{x_\alpha \in \bar{\omega}_\alpha} d_\alpha(x_\alpha).$$

Il reste à trouver $\lambda_{\min}$. Si $q = \varkappa_{-\alpha} = \varkappa_{+\alpha} = 0$, l'opérateur A_α est dégénéré et $\lambda_{\min} = 0$. Autrement dit, en vertu de la remarque 2 du lemme 14, § 2, ch. V, on a

$$(d_\alpha y, y)_{\bar{\omega}_\alpha} \leqslant \max_{\bar{x}_\alpha \in \bar{\omega}_\alpha} w^\alpha(x_\alpha)(A_\alpha y, y)_{\bar{\omega}_\alpha}, \tag{15}$$

où $w^\alpha(x_\alpha)$ est la solution du problème aux limites suivant:

$$\begin{gathered} (a_\alpha w^\alpha_{\bar{x}_\alpha})_{x_\alpha} - 0{,}5 q w^\alpha = -d_\alpha(x_\alpha), \quad h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ \frac{2}{h_\alpha} a_\alpha(h_\alpha) w^\alpha_{x_\alpha} - \left(0{,}5q + \frac{2}{h_\alpha}\varkappa_{-\alpha}\right) w^\alpha = -d_\alpha(0), \quad x_\alpha = 0. \\ -\frac{2}{h_\alpha} a_\alpha(l_\alpha) w^\alpha_{\bar{x}_\alpha} - \left(0{,}5q + \frac{2}{h_\alpha}\varkappa_{+\alpha}\right) w^\alpha = -d_\alpha(l_\alpha), \quad x_\alpha = l_\alpha. \end{gathered} \tag{16}$$

En multipliant (15) par $\hbar_\beta(x_\beta)$ et en sommant en x_β de 0 à l_β, il vient

$$(\mathscr{D}y, y) \leqslant \max_{\bar{x}_\alpha \in \bar{\omega}_\alpha} w^\alpha(x_\alpha)(A_\alpha y, y)$$

et, par conséquent,

$$\lambda_{\min} \geqslant \frac{1}{\max\limits_{x_\alpha \in \bar{\omega}_\alpha} w^\alpha (x_\alpha)}$$

Bref, si $q = \varkappa_{-\alpha} = \varkappa_{+\alpha} = 0$, on a

$$\delta_\alpha = 0, \quad \Delta_\alpha = 2 \max_{x_\alpha \in \bar{\omega}_\alpha} d_\alpha (x_\alpha),$$

autrement dit

$$\delta_\alpha = \frac{1}{\max\limits_{x_\alpha \in \bar{\omega}_\alpha} v^\alpha (x_\alpha)},$$

$$\Delta_\alpha = \left(2 - \frac{1}{\max\limits_{x_\alpha \in \bar{\omega}_\alpha} w^\alpha (x_\alpha)}\right) \max_{x_\alpha \in \bar{\omega}_\alpha} d_\alpha (x_\alpha),$$

où $v^\alpha (x_\alpha)$ et $w^\alpha (x_\alpha)$ sont les solutions des problèmes (13) et (16). Toute l'information à priori nécessaire à la mise en œuvre de la méthode des directions alternées est trouvée. En utilisant les formules du théorème 1, on obtient les paramètres d'itération de la méthode et on est en mesure d'apprécier le nombre exigé d'itérations.

Donnons maintenant les formules de l'algorithme de la méthode des directions alternées pour l'exemple considéré. Compte tenu de la définition des opérateurs A_1, A_2 et du second membre φ, on obtient

$$\omega_{k+1}^{(1)} y_{k+1/2} - \Lambda_1 y_{k+1/2} = \omega_{k+1}^{(2)} y_k + \Lambda_2 y_k + \varphi,$$

$$0 \leqslant x_1 \leqslant l_1, \quad 0 \leqslant x_2 \leqslant l_2,$$

$$\omega_{k+1}^{(2)} y_{k+1} - \Lambda_2 y_{k+1} = \omega_{k+1}^{(2)} y_{k+1/2} + \Lambda_1 y_{k+1/2} + \varphi,$$

$$0 \leqslant x_2 \leqslant l_2, \quad 0 \leqslant x_1 \leqslant l_1.$$

Ici, à la différence du problème de Dirichlet, les problèmes aux limites triponctuels doivent également posséder une solution aux frontières du maillage $\bar{\omega}$, tandis que l'approximation initiale y_0 est une fonction de maille arbitraire donnée sur tout le maillage $\bar{\omega}$.

En profitant des conditions (10), on est en mesure de montrer que pour l'exemple considéré, comme au cas du problème de Dirichlet, pour le nombre d'itérations n se vérifie l'estimation asymptotique suivante en h:

$$n \geqslant n_0 (\varepsilon) = O (\ln |h| \ln \varepsilon), \quad |h|^2 = h_1^2 + h_2^2.$$

Remarquons que toutes les études faites ici conservent leur raison d'être aussi bien dans le cas où ω est un maillage irrégulier quelcon que dans le domaine $\bar{G}$. Il ne faut que remplacer les opérateurs Λ_α introduits ici par des opérateurs associés au maillage irrégulier.

Soulignons une grande importance des hypothèses que q, $\varkappa_{\pm\alpha}$ sont constants et que les coefficients a_α ne dépendent que de x_α. Si l'une au moins de ces hypothèses n'est pas vérifiée, la condition de commutativité des opérateurs A_1 et A_2 ne sera pas alors satisfaite.

En conclusion, notons que la méthode des directions alternées peut être appliquée à la résolution des analogues discrets de l'équation (9) à laquelle sont imposées des conditions aux limites différentes. En particulier, chacun des côtés du rectangle $\bar{G}$ peut être soumis à l'une des conditions aux limites de première, de seconde ou de troisième espèce avec constantes $\varkappa_{\pm\alpha}$.

3. Problème discret de Dirichlet d'ordre élevé de précision. Examinons encore un exemple d'application de la méthode des directions alternées. Soit un maillage rectangle $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leqslant i \leqslant N_1, 0 \leqslant j \leqslant N_2, h_\alpha = l_\alpha/N_\alpha, \alpha = 1, 2\}$, introduit sur le rectangle $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha, \alpha = 1, 2\}$, sur lequel il s'agit de trouver la solution du problème discret de Dirichlet d'ordre de précision élevé pour l'équation de Poisson

$$\Lambda y = (\Lambda_1 + \Lambda_2) + (\varkappa_1 + \varkappa_2) \Lambda_1\Lambda_2) y = -\varphi (x), \quad x \in \omega,$$
$$y (x) = g (x), \quad x \in \gamma, \tag{17}$$

où $\Lambda_\alpha y = y_{\bar{x}_\alpha x_\alpha}$, $\varkappa_\alpha = h_\alpha^2/12$, $\alpha = 1, 2$.

Ici

$$\varphi = \tilde{f} + \varkappa_1\Lambda_1\tilde{f} + \varkappa_2\Lambda_2\tilde{f},$$

où $\tilde{f} (x)$ est le second membre de l'équation différentielle de départ

$$Lu = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -\tilde{f}(x), \quad x \in G, \quad u (x) = g (x), \quad x \in \Gamma.$$

Le schéma aux différences (17), avec le choix indiqué de $\varphi (x)$, possède la précision $O (| h^4 |)$, $| h |^2 = h_1^2 + h_2^2$, tandis que sur le maillage carré ($h_1 = h_2 = h$), avec le choix adéquat de $\varphi (x)$,

$$\varphi = \tilde{f} + \frac{h^2}{12} (\Lambda_1 + \Lambda_2) \tilde{f} + \frac{h^4}{360} (\Lambda_1^2 + 4\Lambda_1\Lambda_2 + \Lambda_2^2) \tilde{f},$$

il possède la précision $O (h^6)$.

En introduisant les opérateurs $A_\alpha y = -\Lambda_\alpha \mathring{y}$, où $y \in H$, $\mathring{y} \in \mathring{H}$ et H, espace des fonctions de mailles données sur ω avec produit scalaire

$$(u, v) = \sum_{x \in \omega} u (x) v (x) h_1 h_2,$$

$\overset{\circ}{H}$ étant l'ensemble des fonctions de mailles s'annulant sur γ, écrivons (17) sous la forme opératorielle

$$Au = f, \tag{18}$$

où $A = A_1 + A_2 - (\varkappa_1 + \varkappa_2) A_1 A_2$.

Comme il a été montré à maintes occasions, les opérateurs A_1 et A_2 possèdent les propriétés suivantes: A_1 et A_2 sont autoadjoints dans H et permutables

$$A_\alpha = A_\alpha^*, \ \alpha = 1, 2, \ A_1 A_2 = A_2 A_1, \tag{19}$$

l'opérateur A_α possédant les bornes δ_α et Δ_α, où

$$\delta_\alpha = \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2l_\alpha}, \quad \Delta_\alpha = \frac{4}{h_\alpha^2} \cos^2 \frac{\pi h_\alpha}{2l_\alpha},$$
$$\delta_\alpha E \leqslant A_\alpha \leqslant \Delta_\alpha E, \ \delta_\alpha > 0, \ \alpha = 1, 2. \tag{20}$$

On a le

Lemme 1. *Si les conditions* (19), (20) *sont remplies et* $\varkappa_\alpha \Delta_\alpha < 1$, *les opérateurs*

$$\bar{A}_\alpha = (E - \varkappa_\alpha A_\alpha)^{-1} A_\alpha, \quad \alpha = 1, 2, \tag{21}$$

sont autoadjoints dans H, permutables et possèdent des bornes $\bar{\delta}_\alpha$ *et* $\bar{\Delta}_\alpha$, *c'est-à-dire*

$$\bar{\delta}_\alpha E \leqslant \bar{A}_\alpha \leqslant \bar{\Delta}_\alpha E, \ \bar{\delta}_\alpha > 0, \ \alpha = 1, 2,$$

où $\bar{\delta}_\alpha$ *et* $\bar{\Delta}_\alpha$ *se définissent par les formules*

$$\bar{\delta}_\alpha = \frac{\delta_\alpha}{1 - \varkappa_\alpha \delta_\alpha}, \quad \bar{\Delta}_\alpha = \frac{\Delta_\alpha}{1 - \varkappa_\alpha \Delta_\alpha}. \tag{22}$$

En effet, l'existence de l'opérateur $\bar{A}_\alpha$ s'ensuit du fait que l'opérateur $E - \varkappa_\alpha A_\alpha$ est défini positif au cas où la condition $\varkappa_\alpha \Delta_\alpha < 1$ est satisfaite. Ensuite, en représentant $\bar{A}_\alpha$ sous la forme de $\bar{A}_\alpha = (A_\alpha^{-1} - \varkappa_\alpha E)^{-1}$ et compte tenu de ce que les opérateurs A_α, A_α^{-1} et $A_\alpha^{-1} - \varkappa_\alpha E$ sont autoadjoints, il vient

$$\left(\frac{1}{\Delta_\alpha} - \varkappa_\alpha\right) E \leqslant (A_\alpha^{-1} - \varkappa_\alpha E) \leqslant \left(\frac{1}{\delta_\alpha} - \varkappa_\alpha\right) E.$$

Il s'ensuit la proposition du lemme. La permutabilité des opérateurs $\bar{A}_1$ et $\bar{A}_2$ est la conséquence de la permutabilité de A_1 et A_2. Le lemme est démontré.

Pour l'exemple considéré les conditions du lemme 1 sont satisfaites vu que $\varkappa_\alpha \Delta_\alpha < 1/3$.

Transformons à présent l'équation (18). Pour ce faire, écrivons (18) sous forme

$$A_1 (E - \varkappa_2 A_2) u + A_2 (E - \varkappa_1 A_1) u = f. \tag{23}$$

En appliquant à (23) l'opérateur $(E - \varkappa_1 A_1)^{-1} (E - \varkappa_2 A_2)^{-1}$ et compte tenu de la permutabilité de tous les opérateurs, on obtient de (23) l'équation équivalente à (18)

$$\bar{A} u = (\bar{A}_1 + \bar{A}_2) u = \bar{f}, \tag{24}$$

où $\bar{A}_1$ et $\bar{A}_2$ sont définis dans (21), tandis que $\bar{f} = (E - \varkappa_1 A_1)^{-1} \times \times (E - \varkappa_2 A_2)^{-1} f$. Bref, la résolution de l'équation (18) est réduite à la résolution de l'équation (24) aux opérateurs $\bar{A}_1$ et $\bar{A}_2$ autoadjoints et permutables, dont les bornes sont définies dans (22).

Pour obtenir la solution approchée de l'équation (24), recourons à la méthode des directions alternées

$$\bar{B}_{k+1} \frac{y_{k+1} - y_k}{\tau_{k+1}} + \bar{A} y_k = \bar{f}, \quad k = 0, 1, \ldots, \quad y_0 \in H, \tag{25}$$

où

$$\bar{B}_k = (\omega_k^{(1)} E + \bar{A}_1)(\omega_k^{(2)} E + \bar{A}_2), \quad \tau_k = \omega_k^{(1)} + \omega_k^{(2)}.$$

Les paramètres d'itération $\omega_k^{(1)}$ et $\omega_k^{(2)}$ s'obtiennent suivant les formules du théorème 1, où δ_α et Δ_α sont remplacés par $\bar{\delta}_\alpha$ et $\bar{\Delta}_\alpha$. Toutes les conditions nécessaires d'applicabilité de la méthode des directions alternées sont dans ce cas satisfaites.

Il reste à examiner l'algorithme mettant en œuvre la méthode itérative (25). Récrivons (25) sous la forme

$$(\omega_{k+1}^{(1)} E + \bar{A}_1)(\omega_{k+1}^{(2)} E + \bar{A}_2) y_{k+1} = \\ = (\omega_{k+1}^{(2)} E - \bar{A}_1)(\omega_{k+1}^{(1)} E - \bar{A}_2) y_k + \tau_{k+1} \bar{f}. \tag{26}$$

On a montré au point 4, § 1, que les paramètres d'itération $\omega_k^{(1)}$ et $\omega_k^{(2)}$ satisfont pour tout k aux inégalités $\bar{\delta}_2 \leqslant \omega_k^{(1)} \leqslant \bar{\Delta}_2$, $\bar{\delta}_1 \leqslant \omega_k^{(2)} \leqslant \bar{\Delta}_1$.

Vu que pour l'exemple considéré $\bar{\delta}_\alpha > 0$, en divisant les deux membres de (26) par $\omega_{k+1}^{(1)} \omega_{k+1}^{(2)}$ et en posant $\tau_{k+1}^{(1)} = 1/\omega_{k+1}^{(1)}$, $\tau_{k+1}^{(2)} = = 1/\omega_{k+1}^{(2)}$, on obtient

$$(E + \tau_{k+1}^{(1)} \bar{A}_1)(E + \tau_{k+1}^{(2)} \bar{A}_2) y_{k+1} = \\ = (E - \tau_{k+1}^{(2)} \bar{A}_1)(E - \tau_{k+1}^{(1)} \bar{A}_2) + (\tau_{k+1}^{(1)} + \tau_{k+1}^{(2)}) \bar{f}.$$

Appliquons aux deux membres de cette égalité l'opérateur

$$(E - \varkappa_1 A_1)(E - \varkappa_2 A_2)$$

et tenons compte de ce que tous les opérateurs sont permutables et que

$$(E - \varkappa_\alpha A_\alpha)(E + \tau_{k+1}^{(\alpha)} \bar{A}_\alpha) = E + (\tau_{k+1}^{(\alpha)} - \varkappa_\alpha) A_\alpha,$$

$$(E - \varkappa_\alpha A_\alpha)(E - \tau_{k+1}^{(\beta)} \bar{A}_\alpha) = E - (\tau_{k+1}^{(\beta)} + \varkappa_\alpha) A_\alpha,$$

$$\beta = 3 - \alpha, \quad \alpha = 1, 2.$$

Finalement, on obtient

$$(E+(\tau_{k+1}^{(1)}-\varkappa_1)A_1)(E+(\tau_{k+1}^{(2)}-\varkappa_2)A_2)y_{k+1}=$$
$$=(E-(\tau_{k+1}^{(2)}+\varkappa_1)A_1)(E-(\tau_{k+1}^{(1)}+\varkappa_2)A_2)y_k+(\tau_{k+1}^{(1)}+\tau_{k+1}^{(2)})f. \quad (27)$$

Le schéma itératif (27) est équivalent au schéma suivant:

$$(E+(\tau_{k+1}^{(1)}-\varkappa_1)A_1)y_{k+1/2}=(E-(\tau_{k+1}^{(1)}+\varkappa_2)A_2)y_k+(\tau_{k+1}^{(1)}-\varkappa_1)f, \quad (28)$$

$$E+(\tau_{k+1}^{(2)}-\varkappa_2)A_2)y_{k+1}=(E-(\tau_{k+1}^{(2)}+\varkappa_1)A_1)y_{k+1/2}+(\tau_{k+1}^{(2)}+\varkappa_1)f. \quad (29)$$

L'équivalence de (27) et de (28), (29) se démontre de la façon suivante. En multipliant (28) par $\tau_{k+1}^{(2)}+\varkappa_1$, (29) par $-(\tau_{k+1}^{(1)}-\varkappa_1)$ et en additionnant les résultats, il vient

$$(\tau_{k+1}^{(1)}+\tau_{k+1}^{(2)})y_{k+1/2}=(\tau_{k+1}^{(1)}-\varkappa_1)(E+(\tau_{k+1}^{(2)}-\varkappa_2)A_2)y_{k+1}+$$
$$+(\tau_{k+1}^{(2)}+\varkappa_1)(E-(\tau_{k+1}^{(1)}+\varkappa_2)A_2)y_k. \quad (30)$$

En portant (30) dans (28), on obtient, après des transformations fort simples, (27). La marche inverse des raisonnements est évidente.

Compte tenu de la définition des opérateurs A_1 et A_2, le schéma (28), (29) peut être écrit sous forme d'un schéma aux différences ordinaire

$$(E-(\tau_{k+1}^{(1)}-\varkappa_1)\Lambda_1)y_{k+1/2}=(E+(\tau_{k+1}^{(1)}+\varkappa_2)\Lambda_2)y_k+(\tau_{k+1}^{(1)}-\varkappa_1)\varphi \quad (31)$$

pour $h_1\leqslant x_1\leqslant l_1-h_1$,

$$y_{k+1/2}=g(x)+(\varkappa_1+\varkappa_2)\Lambda_2 g(x),\ x_1=0,\ l_1.$$

Le problème aux limites (31) doit être résolu de proche en proche pour $h_2\leqslant x_2\leqslant l_2-h_2$. On trouvera ainsi $y_{k+1/2}$ pour $0\leqslant x_1\leqslant l_1$, $h_2\leqslant x_2\leqslant l_2-h_2$. Ensuite,

$$(E-(\tau_{k+1}^{(2)}-\varkappa_2)\Lambda_2)y_{k+1}=(E+(\tau_{k+1}^{(2)}+\varkappa_1)\Lambda_1)y_{k+1/2}+(\tau_{k+1}^{(2)}+\varkappa_1)\varphi \quad (32)$$

pour $h_2\leqslant x_2\leqslant l_2-h_2$,

$$y_{k+1}=g(x),\quad x_2=0,\ l_2.$$

Le problème aux limites (32) doit être résolu de proche en proche pour $h_1\leqslant x_1\leqslant l_1-h_1$. Finalement on obtient y_{k+1}.

Si l'on confronte les nombres d'itérations de la méthode des directions alternées du schéma aux différences de second ordre de précision, étudié au point 1, § 2, et du schéma d'ordre élevé de précision, le nombre d'itérations dans ce dernier cas sera quelque peu supérieur. Dans le cas particulier de $l_1=l_2=l$, $N_1=N_2=N$, si

$N = 10$, l'accroissement est de 1 %, tandis que pour $N = 100$, il est de 4%. Le volume des calculs nécessités par chaque itération est, dans les deux schémas, pratiquement le même, quant à la différence entre les nombres d'itérations, elle est de peu d'importance. Etant donné que le schéma d'ordre élevé de précision permet d'utiliser un maillage plus grossier pour obtenir la précision imposée à la solution du problème différentiel, son application s'avère particulièrement rentable au cas où la solution du problème différentiel est suffisamment lisse.

Rappelons qu'au § 3, ch. III, pour résoudre le problème (17), on a eu recourt à une méthode directe, la méthode de réduction. Pour le schéma de second ordre de précision, comme pour le schéma étudié ici, la méthode directe exigera un nombre d'opérations arithmétiques moindre que la méthode des directions alternées à paramètres optimaux.

§ 3. Méthode des directions alternées dans le cas général

1. Cas des opérateurs non permutables. Supposons qu'il s'agit de trouver la solution de l'équation linéaire opératorielle

$$Au = f \tag{1}$$

avec l'opérateur A non dégénéré qu'on représentera sous forme d'une somme de deux opérateurs autoadjoints non permutables A_1 et A_2 aux bornes δ_1, Δ_1 et δ_2, Δ_2:

$$A_\alpha = A_\alpha^*,\ \delta_\alpha E \leqslant A_\alpha < \Delta_\alpha E,\quad \alpha = 1,\ 2,\ \delta_1 + \delta_2 > 0, \qquad A = A_1 + A_2. \tag{2}$$

Pour pouvoir résoudre de façon approchée l'équation (1), examinons le schéma à deux couches de la méthode des directions alternées à deux paramètres d'itération $\omega^{(1)}$ et $\omega^{(2)}$:

$$B\frac{y_{k+1}-y_k}{\tau} + Ay_k = f,\ k = 0,\ 1,\ \ldots,\ y_0 \in H, \qquad B = (\omega^{(1)}E + A_1)(\omega^{(2)}E + A_2),\ \tau = \omega^{(1)} + \omega^{(2)}. \tag{3}$$

Les paramètres d'itération et l'opérateur B ne dépendent pas ici du numéro d'itération k.

Comme au cas d'opérateurs permutables A_1 et A_2, l'approximation itérative y_{k+1} pour le schéma (3) peut être obtenue au moyen de l'algorithme suivant:

$$(\omega^{(1)}E + A_1)\, y_{k+1/2} = (\omega^{(1)}E - A_2)\, y_k + f,$$
$$(\omega^{(2)}E + A_2)\, y_{k+1} = (\omega^{(2)}E - A_1)\, y_{k+1/2} + f,\ k = 0,\ 1,\ \ldots$$

Etudions la convergence du schéma itératif (3) et cherchons les valeurs optimales des paramètres $\omega^{(1)}$ et $\omega^{(2)}$. En admettant que

l'opérateur $\omega^{(2)}E + A_2$ est non dégénéré, étudions la convergence de (3) dans l'espace énergétique H_D, où $D = (\omega^{(2)}E + A_2)^2$. En vertu de (2), l'opérateur D est autoadjoint dans H et, d'autre part, de l'hypothèse posée plus haut il s'ensuit que D est défini positif dans H.

Pour l'erreur $z_k = y_k - u$, on obtient de (3) l'équation homogène

$$z_{k+1} = Sz_k, \quad k = 0, 1, \ldots, \quad z_0 = y_0 - u, \tag{4}$$
$$S = (\omega^{(2)}E + A_2)^{-1} (\omega^{(1)}E + A_1)^{-1} (\omega^{(2)}E - A_1) (\omega^{(1)}E - A_2).$$

Passons dans (4) au problème de l'erreur équivalente

$$x_k = (\omega^{(2)}E + A_2) z_k. \tag{5}$$

Il vient alors

$$\begin{gathered} x_{k+1} = \bar{S}x_k, \quad k = 0, 1, \ldots, \quad \bar{S} = \bar{S}_1\bar{S}_2, \\ \bar{S}_1 = (\omega^{(1)}E + A_1)^{-1} (\omega^{(2)}E - A_1), \\ \bar{S}_2 = (\omega^{(2)}E + A_2)^{-1} (\omega^{(1)}E - A_2). \end{gathered} \tag{6}$$

Vu qu'en vertu de la substitution réalisée (5) on aboutit à l'égalité $\| x_k \| = \| z_k \|_D$, il suffit d'étudier le comportement de la norme de l'erreur équivalente x_k dans l'espace H. De (6) on tire

$$\| x_{k+1} \| \leqslant \| \bar{S} \| \, \| x_k \| \leqslant \| \bar{S}_1 \| \, \| \bar{S}_2 \| \, \| x_k \|, \quad k = 0, 1, \ldots$$

et, par conséquent,

$$\| z_n \|_D = \| x_n \| \leqslant \| \bar{S} \|^n \, \| x_0 \| \leqslant (\| \bar{S}_1 \| \, \| \bar{S}_2 \|)^n \, \| z_0 \|_D. \tag{7}$$

Apprécions la norme des opérateurs $\bar{S}_1$ et $\bar{S}_2$. Supposons que les opérateurs $\omega^{(\alpha)}E + A_\alpha$, $\alpha = 1, 2$, ne sont pas négatifs. Alors sur la base du point 4, § 1, ch. V, en vertu de (2), on obtient

$$\| \bar{S}_1 \| \leqslant \max_{\delta_1 \leqslant x \leqslant \Delta_1} \left| \frac{\omega^{(2)} - x}{\omega^{(1)} + x} \right|, \qquad \| \bar{S}_2 \| \leqslant \max_{\delta_2 \leqslant y \leqslant \Delta_2} \left| \frac{\omega^{(1)} - y}{\omega^{(2)} + y} \right|$$

et, par conséquent,

$$\| \bar{S}_1 \| \, \| \bar{S}_2 \| \leqslant \max_{\substack{\delta_1 \leqslant x \leqslant \Delta_1 \\ \delta_2 \leqslant y \leqslant \Delta_2}} | R_1(x, y) |, \qquad R_1(x, y) = \frac{\omega^{(2)} - x}{\omega^{(1)} + x} \, \frac{\omega^{(1)} - y}{\omega^{(2)} + y}.$$

Compte tenu de l'estimation (7), posons le problème de la recherche des paramètres $\omega^{(1)}$ et $\omega^{(2)}$ à partir de la condition

$$\min_{\omega^{(1)}, \, \omega^{(2)}} \max_{\substack{\delta_1 \leqslant x \leqslant \Delta_1 \\ \delta \leqslant y \leqslant \Delta_2}} | R_1(x, y) |.$$

Ce problème est un cas particulier du problème résolu au § 1 du présent chapitre. A l'aide de la transformation fractionnaire linéaire des variables x et y (voir (15), (21)-(24) au § 1) le problème posé se réduit au problème de la recherche des paramètres $\varkappa^*$ sur la base de la condition

$$\max_{\eta \leqslant u \leqslant 1} \left| \frac{\varkappa^* - u}{\varkappa^* + u} \right| = \min_{\varkappa} \max_{\eta \leqslant u \leqslant 1} \left| \frac{\varkappa - u}{\varkappa + u} \right| = \rho. \tag{8}$$

En outre, les paramètres $\omega^{(1)}$ et $\omega^{(2)}$ s'expriment en fonction de $\varkappa^*$ à l'aide des formules

$$\omega^{(1)}=\frac{r\varkappa^*+s}{1+t\varkappa^*}, \qquad \omega^{(2)}=\frac{r\varkappa^*-s}{1-t\varkappa^*},$$

tandis que pour l'erreur z_n on a l'estimation

$$\| z_n \|_D \leqslant \rho^{2n} \| z_0 \|_D.$$

D'autre part, on a montré au point 4, § 1, que lors du choix optimal de $\varkappa^*$ les opérateurs $\omega^{(\alpha)}E + A_\alpha$ sont définis positifs si $\delta_1 + \delta_2 > 0$. Donc, en vertu de (2), nos hypothèses sur la non-négativité des opérateurs $\omega^{(\alpha)}E + A_\alpha$, $\alpha = 1, 2$, seront à plus forte raison satisfaites.

Donnons la solution du problème (8) indépendamment des résultats acquis au point 4, § 1. Voyons la fonction $\varphi(u) = (\varkappa - u)/(\varkappa + u)$ sur le segment $0 < \eta \leqslant u \leqslant 1$ pour $\varkappa > 0$. Cette fonction décroît de façon monotone en u et, par conséquent,

$$\max_{\eta\leqslant u\leqslant 1} |\varphi(u)| = \max\left(\left|\frac{\eta-\varkappa}{\eta+\varkappa}\right|, \left|\frac{1-\varkappa}{1+\varkappa}\right|\right).$$

De là on obtient sans peine que le minimum en $\varkappa$ de cette expression est atteint pour $\varkappa^*$, qu'on définit sur la base de l'égalité

$$\frac{\varkappa^*-\eta}{\varkappa^*+\eta}=\frac{1-\varkappa^*}{1+\varkappa^*}.$$

Il en résulte que

$$\varkappa^*=\sqrt{\eta},\ \min_{\varkappa}\max_{\eta\leqslant u\leqslant 1}\left|\frac{\varkappa-u}{\varkappa+u}\right|=\rho=\frac{1-\sqrt{\eta}}{1+\sqrt{\eta}}.$$

On a ainsi démontré le théorème 2.

T h é o r è m e 2. *Soient les conditions* (2) *remplies, quant aux paramètres* $\omega^{(1)}$ *et* $\omega^{(2)}$, *ils sont choisis suivant les formules*

$$\omega^{(1)}=\frac{r\sqrt{\eta}+s}{1+t\sqrt{\eta}}, \qquad \omega^{(2)}=\frac{r\sqrt{\eta}-s}{1-t\sqrt{\eta}},$$

où r, s, t *et* η *sont définis dans* (21)-(24), § 1. *La méthode des directions alternées* (3) *converge dans* H_D *et pour l'erreur* z_n *on a l'estimation*

$$\| z_n \|_D\leqslant\rho^{2n} \| z_0 \|_D, \quad \rho=\frac{1-\sqrt{\eta}}{1+\sqrt{\eta}},$$

où $D=(\omega^{(2)}E+A_2)^2$. *Pour le nombre d'itérations* n *se vérifie l'estimation*

$$n=n_0(\varepsilon)=\ln\varepsilon/(2\ln\rho)\approx\ln\frac{1}{\varepsilon}\Big/(4\sqrt{\eta}).$$

2. Problème discret de Dirichlet pour l'équation elliptique à coefficients variables. Examinons l'exemple d'application de la

méthode des directions alternées au cas de non-commutativité. Supposons qu'il s'agit de rechercher sur un maillage rectangulaire $\bar\omega = \{x_{ij} = (ih_1, jh_2) \in \bar G, 0 \leqslant i \leqslant N_1, 0 \leqslant j \leqslant N_2, h_\alpha = l_\alpha/N_\alpha, \alpha = 1, 2\}$, introduit dans le rectangle $\bar G = \{0 \leqslant x_\alpha \leqslant l_\alpha, \alpha = 1, 2\}$, la solution du problème de différences suivant:

$$\Lambda y = \sum_{\alpha=1}^{2} (a_\alpha(x)\, y_{\bar x_\alpha})_{x_\alpha} - q(x)\, y = -\varphi(x), \quad x \in \omega, \tag{9}$$

$$y(x) = g(x), \quad x \in \gamma,$$

où les coefficients du schéma aux différences satisfont aux conditions

$$0 < c_1 \leqslant a_\alpha(x) \leqslant c_2, \quad 0 \leqslant d_1 \leqslant q(x) \leqslant d_2. \tag{10}$$

Dans l'espace H des fonctions de mailles données sur ω avec le produit scalaire $(u, v) = \sum_{x\in\omega} u(x)\, v(x)\, h_1 h_2$ on définira les opérateurs A_1 et A_2 de la façon suivante: $A_\alpha y = -\Lambda_\alpha \overset{\circ}{y} = -(a_\alpha \overset{\circ}{y}_{\bar x_\alpha})_{x_\alpha} + 0{,}5q\overset{\circ}{y}$, $\alpha = 1, 2$, $y \in H$, $\overset{\circ}{y} \in \overset{\circ}{H}$, où, comme habituellement, $\overset{\circ}{y}(x) = 0$ sur γ.

Les opérateurs introduits A_α sont autoadjoints dans H et si $a_\alpha(x)$ ne dépend que de la variable x_α et $q(x)$ est une constante, les opérateurs A_1 et A_2 seront permutables. Dans le cas général, par contre, la permutabilité n'aura pas lieu et, pour résoudre l'équation (1) correspondant au problème de différences (9), on peut recourir à la méthode des directions alternées (3) étudiée au point 1, § 3.

L'algorithme de la méthode acquiert une forme simple

$$\omega^{(1)} y_{k+1/2} - \Lambda_1 y_{k+1/2} = \omega^{(1)} y_k + \Lambda_2 y_k + \varphi, \quad h_1 \leqslant x_1 \leqslant l_1 - h_1,$$

$$y_{k+1/2}(x) = g(x), \quad x_1 = 0, l_1, \quad h_2 \leqslant x_2 \leqslant l_2 - h_2,$$

$$\omega^{(2)} y_{k+1} (-\Lambda_2 y_{k+1} = \omega^{(2)} y_{k+1/2} + \Lambda_1 y_{k+1/2} + \varphi, \quad h_2 \leqslant x_2 \leqslant l_2 - h_2,$$

$$y_{k+1}(x) = g(x), \quad x_2 = 0, l_2, \quad h_1 \leqslant x_1 \leqslant l_1 - h_1.$$

Il reste à trouver les bornes δ_α et Δ_α des opérateurs A_α, $\alpha = 1, 2$. Les conditions (10) étant remplies, on obtient sur la base du lemme 14, § 2, ch. V

$$(y^2, 1)_{\omega_\alpha} \leqslant \varkappa_\alpha(x_\beta)\, (A_\alpha y, y)_{\omega_\alpha}, \quad \beta = 3 - \alpha, \quad \alpha = 1, 2, \tag{11}$$

où $\varkappa_\alpha(x_\beta) = \max_{x_\alpha \in \omega_\alpha} v^\alpha(x)$, tandis que $v^\alpha(x)$ est la solution du problème aux limites triponctuel suivant:

$$(a_\alpha v^\alpha_{\bar x_\alpha})_{x_\alpha} - 0{,}5 q v^\alpha = -1, \quad h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha,$$

$$v^\alpha(x) = 0, \quad x_\alpha = 0, l_\alpha, \quad h_\beta \leqslant x_\beta \leqslant l_\beta - h_\beta.$$

Le produit scalaire en ω_α se définit de la façon suivante:

$$(u, v)_{\omega_\alpha} = \sum_{x_\alpha=h_\alpha}^{l_\alpha-h_\alpha} u(x)\, v(x)\, h_\alpha = \sum_{x_\alpha \in \omega_\alpha} u(x)\, v(x)\, h_\alpha.$$

En multipliant (11) par h_β et en sommant en x_β, il vient

$$\left(\frac{1}{\varkappa_\alpha} y^2,\ 1\right) \leqslant (A_\alpha y,\ y), \qquad \alpha = 1,\ 2.$$

Par conséquent, en guise de δ_α il est possible de prendre

$$\delta_\alpha = \min_{h_\beta \leqslant x_\beta \leqslant l_\beta - h_\beta} \frac{1}{\varkappa_\alpha(x_\beta)}, \qquad \beta = 3-\alpha,\ \alpha = 1,\ 2.$$

Cherchons maintenant Δ_α. Opérons de façon analogue qu'au point 2, § 2. Désignons par $\mathscr{D}$ la partie diagonale de la matrice $\mathscr{A}_\alpha$ correspondant à l'opérateur A_α:

$$\mathscr{D} y = d_\alpha(x)\, y,$$

$$d_\alpha(x) = \begin{cases} 0{,}5q(x) + \dfrac{1}{h_\alpha^2}\left(a_\alpha(x_\alpha,\ x_\beta) + a_\alpha(x_\alpha + h_\alpha,\ x_\beta)\right), \\ h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ h_\beta \leqslant x_\beta \leqslant l_\beta - h_\beta. \end{cases}$$

On a alors l'inégalité

$$(A_\alpha y,\ y) \leqslant (2-\lambda_{\min})\,(\mathscr{D} y,\ y) \leqslant (2-\lambda_{\min}) \max_{x \in \omega} d_\alpha(x)\,(y,\ y),$$

où $\lambda_{\min}$ est une constante de l'inégalité opératorielle $\lambda_{\min}\mathscr{D} \leqslant A_\alpha$.

Cherchons $\lambda_{\min}$. A partir du lemme 14, § 2, ch. V, on obtient

$$(d_\alpha y^2,\ 1)_{\omega_\alpha} \leqslant \rho_\alpha(x_\beta)\,(A_\alpha y,\ y)_{\omega_\alpha}, \tag{12}$$

où $\rho_\alpha(x_\beta) = \max\limits_{x_\alpha \in \omega_\alpha} w^\alpha(x)$, tandis que $w^\alpha(x)$ est la solution du problème aux limites triponctuel suivant:

$$(a_\alpha w^\alpha_{\bar{x}_\alpha})_{x_\alpha} - 0{,}5\, q w^\alpha = -d_\alpha(x), \qquad h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha,$$
$$w^\alpha(x) = 0, \quad x_\alpha = 0, \quad l_\alpha, \quad h_\beta \leqslant x_\beta \leqslant l_\beta - h_\beta.$$

En multipliant (12) par h_β et en sommant en ω_β, il vient

$$\left(\frac{d_\alpha}{\rho_\alpha} y^2,\ 1\right) \leqslant (A_\alpha y,\ y), \quad \alpha = 1,\ 2.$$

Par conséquent, en guise de $\lambda_{\min}$ on peut prendre

$$\lambda_{\min} = \min_{h_\beta \leqslant x_\beta \leqslant l_\beta - h_\beta} \frac{1}{\rho_\alpha(x_\beta)},$$

et Δ_α a donc pour expression

$$\Delta_\alpha = \left(2 - \frac{1}{\max\limits_{x_\beta} \rho_\alpha(x_\beta)}\right) \max_{x \in \omega} d_\alpha(x), \quad \alpha = 1,\ 2.$$

Bref, l'information à priori exigée pour la mise en œuvre de la méthode des directions alternées est obtenue. En utilisant les conditions (10), on peut montrer que la grandeur η déterminant la vitesse de convergence de la méthode pour l'exemple considéré vaut $O(|h|^2)$, où $|h|^2 = h_1^2 + h_2^2$. Aussi, en vertu du théorème 2, se vérifie-t-elle l'estimation suivante pour le nombre d'itérations

$$n = O\left(\frac{1}{|h|} \ln \frac{1}{\varepsilon}\right).$$

Voyons un problème modèle. Supposons que le schéma aux différences (9) est donné sur un maillage carré dans un carré unitaire ($N_1 = N_2 = N$, $l_1 = l_2 = 1$). Les coefficients $a_1(x)$, $a_2(x)$ et $q(x)$ seront choisis de la façon suivante:

$$a_1(x) = 1 + c\,[(x_1 - 0{,}5)^2 + (x_2 - 0{,}5)^2],$$
$$a_2(x) = 1 + c\,[0{,}5 - (x_1 - 0{,}5)^2 - (x_2 - 0{,}5)^2],$$
$$q(x) \equiv 0, \quad c > 0.$$

Dans ce cas dans les inégalités (10) $c_1 = 1$, $c_2 = 1 + 0{,}5c$, $d_1 = d_2 = 0$; en variant le paramètre c, on obtient les coefficients du schéma aux différences (9) aux caractéristiques extrémales variées.

Donnons le nombre d'itérations de la méthode considérée des directions alternées en fonction du rapport c_2/c_1 et du nombre de nœuds N suivant une direction pour $\varepsilon = 10^{-4}$.

Tableau 12

c_2/c_1	$N = 32$	$N = 64$	$N = 128$
2	65	132	264
8	90	187	380
32	110	233	482
128	122	264	556
512	128	282	603

Comparons cette méthode avec la méthode de surrelaxation (voir § 2, ch. IX), la méthode triangulaire alternée (voir § 2, ch. X) et la méthode implicite de Tchébychev (voir point 3, § 2, ch. VI). En ce qui concerne le nombre d'itérations, la méthode étudiée des directions alternées n'est pas à la hauteur de la méthode de surrelaxation et de la méthode triangulaire alternée, mais elle est supérieure à la méthode implicite de Tchébychev de 1,5 à 2 fois. Cependant, quant au volume des calculs, la méthode des directions alternées est inférieure à la méthode implicite de Tchébychev.

CHAPITRE XII

MÉTHODES DE RÉSOLUTION DES ÉQUATIONS À ORÉRATEURS DÉGÉNÉRÉS DE SIGNES INDÉTERMINÉS

On étudie dans ce chapitre les méthodes directes et itératives permettant de résoudre les équations possédant un opérateur non dégénéré et de signe indéterminé, un opérateur complexe, ainsi qu'un opérateur dégénéré. Le § 1 est consacré aux équations à opérateur de signe indéterminé résolues à l'aide de la méthode à paramètres de Tchébychev et de la méthode du type variationnel. Dans le § 2, on construit pour les équations à opérateur complexe la méthode itérative simple et la méthode des directions alternées avec paramètres d'itération complexes. Dans le § 3 on étudie les méthodes itératives générales de résolution des équations à opérateur dégénéré dans le cas, où sur la couche supérieure l'opérateur est non dégénéré. Le § 4 traite de la construction des méthodes directes et itératives spéciales pour des équations à opérateur dégénéré.

§ 1. Equations à opérateur réel de signe indéterminé

1. Schéma itératif. Problème de choix des paramètres d'itération. Soit donnée dans l'espace hilbertien H l'équation

$$Au = f \tag{1}$$

à opérateur A linéaire et non dégénéré. Afin de résoudre l'équation (1), prenons le schéma itératif implicite à deux couches

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}}+Ay_k=f, \quad k=0, 1, \ldots, \tag{2}$$

à opérateur non dégénéré B et $y_0 \in H$ quelconque.

Les schémas itératifs de la forme (2) ont été étudiés dans les ch. VI, VIII, où on a proposé quelques procédés de choix des paramètres d'itération τ_k en fonction des propriétés des opérateurs A, B et D. Rappelons que D est un opérateur autoadjoint et défini positif engendrant l'espace énergétique H_D. On a montré que pour la convergence dans H_D des méthodes itératives considérées il faut que l'opérateur

$$C = D^{-1/2}(DB^{-1}A)D^{-1/2} \tag{3}$$

soit défini positif. Pour des opérateurs D concrets cette exigence aboutit aux conditions suivantes imposées aux opérateurs A et B:

1) l'opérateur A doit être défini positif dans H si $D = A$, B ou $A^*B^{-1}A$;

2) l'opérateur B^*A doit être défini positif dans H si $D = A^*A$ ou B^*B.

Il existe des problèmes pour lesquels ces exigences ne sont pas satisfaites, c'est-à-dire que soit l'opérateur A est de signe non déterminé, soit qu'il s'avère difficile de trouver un opérateur B tel que B^*A devienne un opérateur défini positif. En guise d'exemple de tels problèmes on peut invoquer le problème de Dirichlet pour l'équation de Helmholtz dans un rectangle

$$y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} + m^2 y = 0, \qquad x \in \omega,$$
$$y(x) = g(x), \qquad x \in \gamma,$$

où $m^2 > 0$.

Dans ce paragraphe on construira des méthodes itératives implicites à deux couches pour le cas où l'opérateur C est un opérateur non dégénéré et de signe indéterminé dans H. On n'envisagera ici que les opérateurs C réels, le cas complexe étant renvoyé au § 2.

Passons à la construction des méthodes itératives. Dans l'équation

$$z_{k+1} = (E - \tau_{k+1} B^{-1} A) z_k, \quad k = 0, 1, \ldots, 2n - 1,$$

effectuons la substitution $z_k = D^{-1/2} x_k$ pour l'erreur $z_k = y_k - u$ du schéma itératif (2) et passons à l'équation de l'erreur équivalente x_k:

$$x_{k+1} = (E - \tau_{k+1} C) x_k, \quad k = 0, 1, \ldots, 2n - 1, \tag{4}$$

où l'opérateur C est défini dans (3). Vu que l'opérateur C est de signe indéterminé, il s'avère apparemment que la norme de l'opérateur $E - \tau_{k+1} C$ sera supérieure ou égale à l'unité pour tout τ_{k+1}.

Examinons maintenant l'équation liant les erreurs sur les itérations paires. De (4) il vient

$$x_{2k+2} = (E - \tau_{2k+2} C)(E - \tau_{2k+1} C) x_{2k}, \quad k = 0, 1, \ldots \ldots, n - 1. \tag{5}$$

Si l'on note

$$\omega_{k+1} = -\tau_{2k+2} \tau_{2k+1}, \quad k = 0, 1, \ldots, n - 1, \tag{6}$$

et l'on exige que les paramètres d'itération τ_{2k+2} et τ_{2k+1} satisfassent pour tout k à la relation

$$1/\tau_{2k+2} + 1/\tau_{2k+1} = 2\alpha, \quad k = 0, 1, \ldots, n - 1, \tag{7}$$

où α est une constante, pour le moment, indéterminée, alors (5) peut être écrit sous la forme

$$x_{2k+2} = (E - \omega_{k+1} \bar{C}) x_{2k}, \quad k = 0, 1, \ldots, \quad \bar{C} = C^2 - 2\alpha C. \tag{8}$$

Si ω_{k+1} et α sont obtenus, les paramètres τ_{2k+2} et τ_{2k+1}, en vertu de (6) et (7), peuvent être déterminés à l'aide des formules

$$\begin{gathered}\tau_{2k+1} = -\alpha\omega_{k+1} - \sqrt{\alpha^2\omega_{k+1}^2 + \omega_{k+1}},\\ \tau_{2k+2} = -\alpha\omega_{k+1} + \sqrt{\alpha^2\omega_{k+1}^2 + \omega_{k+1}},\\ k = 0,\ 1,\ \ldots,\ n-1.\end{gathered} \tag{9}$$

De (8) on obtient

$$x_{2n} = \prod_{j=1}^{n} (E - \omega_j \bar{C})\, x_0,$$

$$\| x_{2n} \| \leqslant \| \prod_{j=1}^{n} (E - \omega_j \bar{C}) \| \, \| x_0 \|. \tag{10}$$

Vu que l'opérateur $\bar{C}$ dépend de α, l'exigence pour l'opérateur $\bar{C}$ d'être défini positif sera l'une des conditions à laquelle est soumis le choix du paramètre α. En outre, il s'ensuit de (10) que les paramètres ω_j, $1 \leqslant j \leqslant n$, et le paramètre α doivent être choisis sur la base de la condition du minimum de la norme de l'opérateur résolvant $\prod_{j=1}^{n} (E = \omega_j \bar{C})$.

Ce problème du meilleur choix des paramètres d'itération ω_j et α et, partant, des paramètres τ_k pour le schéma (2) sera résolu plus loin. Etablissons d'abord la liaison du procédé proposé de la construction de la méthode itérative avec le procédé basé sur la transformation de Gauss au cas d'un opérateur C autoadjoint.

Notons que la substitution $u = D^{-1/2}x$, $f = BD^{-1/2}\varphi$ autorise d'écrire l'équation (1) sous la forme suivante:

$$Cx = \varphi, \tag{11}$$

où l'opérateur C est défini dans (3). En utilisant (11), il vient

$$\bar{C}x = C^2x - 2\alpha Cx = (C - 2\alpha E)\,\varphi = \bar{\varphi}. \tag{12}$$

Ensuite, en posant $v_k = D^{1/2}y_k$, où y_k est l'approximation itérative dans le schéma (2), on obtient sans peine

$$x_k = D^{1/2}z_k = D^{1/2}y_k - D^{1/2}u = v_k - x.$$

En portant x_k dans (8) et compte tenu de (12), on aboutit au schéma itératif

$$\frac{v_{2k+2} - v_{2k}}{\omega_{k+1}} + \bar{C}v_{2k} = \bar{\varphi}, \qquad k = 0,\ 1,\ \ldots \tag{13}$$

Bref, le schéma (13) est un schéma explicite à deux couches pour l'équation transformée (12).

Soit $C = C^*$. Rappelons que dans ce cas la première transformation de Gauss consiste dans le passage de l'équation (11) à l'équation $\bar{\bar{C}}x = C^2x = C\varphi = \bar{\bar{\varphi}}$. Vu que C est un opérateur non dégénéré,

l'opérateur C^2 sera défini positif dans H. Aussi la transformation mentionnée nous ramène-t-elle au cas d'un opérateur de signe déterminé. Pour résoudre une équation munie d'un tel opérateur, on peut recourir à un schéma à deux couches de la forme (13), en y substituant $\bar{\bar{C}}$ à $\bar{C}$ et $\bar{\bar{\varphi}}$ à $\bar{\varphi}$. Il va de soi qu'une telle méthode n'est qu'un cas particulier (pour $\alpha = 0$) de la méthode étudiée.

2. Transformation de l'opérateur au cas où ce dernier est autoadjoint. Supposons que l'opérateur C est autoadjoint dans H. Dans ce cas l'opérateur $\bar{C} = C^2 - \alpha C$ est également autoadjoint dans H. Notre objectif le plus immédiat est de choisir le paramètre α de telle sorte que l'opérateur $\bar{C}$ soit défini positif et de trouver les bornes $\gamma_1 = \gamma_1(\alpha)$ et $\gamma_2 = \gamma_2(\alpha)$ de cet opérateur, c'est-à-dire les grandeurs des inégalités

$$\gamma_1 E \leqslant \bar{C} \leqslant \gamma_2 E, \quad \gamma_1 > 0. \tag{14}$$

Si la valeur indiquée de α existe, alors, en vertu de l'estimation

$$\left\| \prod_{j=1}^{n} (E - \omega_j \bar{C}) \right\| \leqslant \max_{\gamma_1 \leqslant t \leqslant \gamma_2} \left| \prod_{j=1}^{n} (1 - \omega_j t) \right|,$$

le problème de l'obtention des paramètres ω_j, $j = 1, 2, \ldots, n$ se réduit à la construction du polynôme $P_n(t)$ de degré n normé par la condition $P_n(0) = 1$ qui s'écarte le moins de zéro sur le segment $[\gamma_1, \gamma_2]$ du demi-axe positif. Ce problème a déjà été étudié au ch. VI lors de la construction de la méthode de Tchébychev. La solution prend la forme

$$\omega_k = \frac{\omega_0}{1 + \rho_0 \mu_k}, \quad \mu_k \in \mathfrak{M}_n^* \left\{ -\cos \frac{2i-1}{2n} \pi, \ i = 1, 2, \ldots, n \right\},$$

où $k = 1, 2, \ldots, n$,

$$\omega_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

De plus, en vertu de (10), pour l'erreur x_{2n} se vérifie l'estimation

$$\| x_{2n} \| \leqslant q_n \| x_0 \|, \quad q_n = 2\rho_1^n/(1 + \rho_1^{2n}).$$

Il s'ensuit de là que le choix du paramètre α doit se plier à la condition du maximum du rapport γ_1/γ_2.

Cherchons la valeur optimale du paramètre α. Supposons que les valeurs propres μ de l'opérateur C se trouvent sur les segments $[\mathring{\gamma}_1, \mathring{\gamma}_2]$ et $[\mathring{\gamma}_3, \mathring{\gamma}_4]$. Vu que l'opérateur C est de signe indéterminé et est non dégénéré, on a

$$\mathring{\gamma}_1 \leqslant \mathring{\gamma}_2 < 0 < \mathring{\gamma}_3 \leqslant \mathring{\gamma}_4. \tag{15}$$

Cherchons les valeurs propres λ de l'opérateur $\bar{C} = C^2 - 2\alpha C$. On voit sans peine que les valeurs propres des opérateurs $\bar{C}$ et C sont

liées par la relation

$$\lambda = \mu^2 - 2\alpha\mu, \quad \mu \in \Omega, \tag{16}$$

où Ω est constitué de deux segments $[\mathring{\gamma}_1, \mathring{\gamma}_2]$ et $[\mathring{\gamma}_3, \mathring{\gamma}_4]$.

Cherchons d'abord les limites sur α qui garantissent la positivité des valeurs propres de λ, c'est-à-dire que l'opérateur $\overline{C}$ est défini positif. En analysant l'inégalité $\mu^2 - 2\alpha\mu > 0$, on trouve qu'elle a lieu pour μ variant à l'extérieur de l'intervalle $[0, 2\alpha]$. Aussi cette inégalité sera-t-elle remplie pour $\mu \in \Omega$, si α satisfait à la condition

$$\mathring{\gamma}_2 < 2\alpha < \mathring{\gamma}_3. \tag{17}$$

Posons que (17) est vérifié. De (16) on obtient que la transformation $\lambda = \lambda(\mu) = \mu^2 - 2\alpha\mu$ est l'application du segment $[\mathring{\gamma}_1, \mathring{\gamma}_2]$ sur le segment $[\lambda_2, \lambda_1]$, et du segment $[\mathring{\gamma}_3, \mathring{\gamma}_4]$ sur le segment $[\lambda_3, \lambda_4]$, où $\lambda_i = \lambda(\mathring{\gamma}_i)$, $1 \leqslant i \leqslant 4$. De cette façon, toutes les valeurs propres de l'opérateur $\overline{C}$ sont positives et se disposent sur les segments $[\lambda_2, \lambda_1] \cup [\lambda_3, \lambda_4]$. Aussi dans les inégalités (14) faut-il poser

$$\gamma_1 = \min(\lambda_2, \lambda_3), \quad \gamma_2 = \max(\lambda_1, \lambda_4). \tag{18}$$

Choisissons maintenant $2\alpha \in (\mathring{\gamma}_2, \mathring{\gamma}_3)$ à partir de la condition du maximum du rapport γ_1/γ_2. De (18), il vient

$$\gamma_1 = \begin{cases} \lambda_2 = \mathring{\gamma}_2(\mathring{\gamma}_2 - 2\alpha), & \mathring{\gamma}_2 < 2\alpha \leqslant \mathring{\gamma}_2 + \mathring{\gamma}_3, \\ \lambda_3 = \mathring{\gamma}_3(\mathring{\gamma}_3 - 2\alpha), & \mathring{\gamma}_2 + \mathring{\gamma}_3 \leqslant 2\alpha < \mathring{\gamma}_3, \end{cases}$$

$$\gamma_2 = \begin{cases} \lambda_4 = \mathring{\gamma}_4(\mathring{\gamma}_4 - 2\alpha), & 2\alpha \leqslant \mathring{\gamma}_1 + \mathring{\gamma}_4, \\ \lambda_1 = \mathring{\gamma}_1(\mathring{\gamma}_1 - 2\alpha), & \mathring{\gamma}_1 + \mathring{\gamma}_4 \leqslant 2\alpha. \end{cases}$$

Introduisons les notations suivantes: $\Delta_1 = \mathring{\gamma}_2 - \mathring{\gamma}_1$, $\Delta_2 = \mathring{\gamma}_4 - \mathring{\gamma}_3$ et examinons deux cas.

1) Soit d'abord $\Delta_1 \leqslant \Delta_2$, c'est-à-dire $\mathring{\gamma}_2 + \mathring{\gamma}_3 \leqslant \mathring{\gamma}_1 + \mathring{\gamma}_4$. Dans ce cas pour $\xi = \gamma_1/\gamma_2$ on aboutit à l'expression suivante

$$\xi = \xi(\alpha) = \begin{cases} \dfrac{\mathring{\gamma}_2(\mathring{\gamma}_2 - 2\alpha)}{\mathring{\gamma}_4(\mathring{\gamma}_4 - 2\alpha)}, & \mathring{\gamma}_2 < 2\alpha \leqslant \mathring{\gamma}_2 + \mathring{\gamma}_3, \text{ croît en } \alpha, \\ \dfrac{\mathring{\gamma}_3(\mathring{\gamma}_3 - 2\alpha)}{\mathring{\gamma}_4(\mathring{\gamma}_4 - 2\alpha)}, & \mathring{\gamma}_2 + \mathring{\gamma}_3 \leqslant 2\alpha \leqslant \mathring{\gamma}_1 + \mathring{\gamma}_4, \text{ décroît en } \alpha, \\ \dfrac{\mathring{\gamma}_3(\mathring{\gamma}_3 - 2\alpha)}{\mathring{\gamma}_1(\mathring{\gamma}_1 - 2\alpha)}, & \mathring{\gamma}_1 + \mathring{\gamma}_4 \leqslant 2\alpha, \text{ décroît en } \alpha. \end{cases}$$

Par conséquent, dans ce cas la valeur optimale de α vaut

$$\alpha = \alpha_0 = (\mathring{\gamma}_2 + \mathring{\gamma}_3)/2, \tag{19}$$

la condition (17) étant remplie. Pour $\alpha = \alpha_0$, il vient

$$\gamma_1 = \lambda_2 = \lambda_3 = -\mathring{\gamma}_2\mathring{\gamma}_3, \tag{20}$$

$$\gamma_2 = \lambda_4 = \mathring{\gamma}_4(\Delta_2 - \Delta_1) - \mathring{\gamma}_1\mathring{\gamma}_4 \geqslant \lambda_1. \tag{21}$$

2) Soit maintenant $\Delta_1 \geqslant \Delta_2$, c'est-à-dire $\mathring{\gamma}_2 + \mathring{\gamma}_3 \geqslant \mathring{\gamma}_1 + \mathring{\gamma}_4$. Dans ce cas on a

$$\xi = \xi(\alpha) = \begin{cases} \dfrac{\mathring{\gamma}_2(\mathring{\gamma}_2 - 2\alpha)}{\mathring{\gamma}_4(\mathring{\gamma}_4 - 2\alpha)}, & 2\alpha \leqslant \mathring{\gamma}_1 + \mathring{\gamma}_4, \text{ croît en } \alpha, \\ \dfrac{\mathring{\gamma}_2(\mathring{\gamma}_2 - 2\alpha)}{\mathring{\gamma}_1(\mathring{\gamma}_1 - 2\alpha)}, & \mathring{\gamma}_1 + \mathring{\gamma}_4 \leqslant 2\alpha \leqslant \mathring{\gamma}_2 + \mathring{\gamma}_3, \text{ croît en } \alpha, \\ \dfrac{\mathring{\gamma}_3(\mathring{\gamma}_3 - 2\alpha)}{\mathring{\gamma}_1(\mathring{\gamma}_1 - 2\alpha)}, & \mathring{\gamma}_2 + \mathring{\gamma}_3 \leqslant 2\alpha < \mathring{\gamma}_3, \text{ décroît en } \alpha. \end{cases}$$

Par conséquent, dans ce cas la valeur optimale du paramètre α se détermine suivant la formule (19), la valeur de γ_1 étant donnée dans (20), tandis que

$$\gamma_2 = \lambda_1 = \mathring{\gamma}_1(\Delta_2 - \Delta_1) - \mathring{\gamma}_1\mathring{\gamma}_4 \geqslant \lambda_4. \tag{22}$$

On a ainsi démontré le lemme 1.

L e m m e 1. *Soient les valeurs propres de l'opérateur C comprises sur les segments* $[\mathring{\gamma}_1, \mathring{\gamma}_2]$ et $[\mathring{\gamma}_3, \mathring{\gamma}_4]$, $\mathring{\gamma}_2 < 0 < \mathring{\gamma}_3$. *Dans ce cas pour l'opérateur* $\bar{C} = C^2 - \alpha C$, si $\alpha = \alpha_0 = (\mathring{\gamma}_2 + \mathring{\gamma}_3)/2$, *se vérifient les inégalités*

$$\gamma_1 E \leqslant \bar{C} \leqslant \gamma_2 E, \quad \gamma_1 > 0,$$

où

$$\gamma_1 = -\mathring{\gamma}_2\mathring{\gamma}_3, \ \gamma_2 = \max[\mathring{\gamma}_4(\Delta_2 - \Delta_1), \ \mathring{\gamma}_1(\Delta_2 - \Delta_1)] - \mathring{\gamma}_1\mathring{\gamma}_4.$$

Pour la valeur mentionnée de α *le rapport* γ_1/γ_2 *est maximal.*

Les assertions du lemme s'ensuivent de (19)-(22). Notons que $\alpha_0 = 0$ seulement au cas où $\mathring{\gamma}_2 = -\mathring{\gamma}_3$.

3. Méthode itérative avec paramètres de Tchébychev. On a vu plus haut le schéma itératif à deux couches (2) dont les paramètres τ_k, $k = 1, 2, \ldots, 2n$, sont exprimés en fonction de ω_k, $1 \leqslant k \leqslant n$ et α suivant les formules (9). Les paramètres ω_k sont dans ce cas des paramètres d'itération de la méthode de Tchébychev et se déterminent au moyen des formules correspondantes, quant à l'information à priori exigée dans ce cas ainsi que la valeur optimale du paramètre α, elles sont fournies par le lemme 1.

Notons qu'on a admis que les valeurs propres μ de l'opérateur autoadjoint C appartenaient aux segments $[\mathring{\gamma}_1, \mathring{\gamma}_2]$ et $[\mathring{\gamma}_3, \mathring{\gamma}_4]$. A par-

tir de la définition (3) de l'opérateur C il s'ensuit que les valeurs propres de l'opérateur C sont en même temps des valeurs propres du problème suivant:

$$Au - \mu Bu = 0. \tag{23}$$

Pour s'en convaincre, il suffit de multiplier cette équation à gauche par l'opérateur $D^{1/2}B^{-1}$ et de procéder à une substitution en posant $u = D^{-1/2}v$. Notons que l'opérateur C sera autoadjoint dans H si l'opérateur $DB^{-1}A$ l'est.

Formulons le résultat obtenu sous forme d'un théorème.

Théorème 1. *Soient l'opérateur $DB^{-1}A$ autoadjoint dans H et les valeurs propres du problème* (23) *appartenant aux segments* $[\mathring{\gamma}_1, \mathring{\gamma}_2]$ *et* $[\mathring{\gamma}_3, \mathring{\gamma}_4]$, $\mathring{\gamma}_1 \leqslant \mathring{\gamma}_2 < 0 < \mathring{\gamma}_3 \leqslant \mathring{\gamma}_4$. *Pour le processus d'itération* (2) *aux paramètres*

$$\tau_{2k-1} = -\alpha_0\omega_k - \sqrt{\alpha_0^2\omega_k^2 + \omega_k}, \qquad \tau_{2k} = -\alpha_0\omega_k + \sqrt{\alpha_0^2\omega_k^2 + \omega_k},$$
$$k = 1, 2, \ldots, n,$$

se vérifie l'estimation

$$\| z_{2n} \|_D \leqslant q_n \| z_0 \|_D,$$

où

$$\omega_k = \frac{\omega_0}{1+\rho_0\mu_k}, \quad \mu_k \in \mathfrak{M}_n^* = \left\{ -\cos\frac{(2i-1)\pi}{2n}, \ 1 \leqslant i \leqslant n \right\}, \quad 1 \leqslant k \leqslant n,$$
$$\omega_0 = \frac{2}{\gamma_1+\gamma_2}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad q_n = \frac{2\rho_1^n}{1+\rho_1^{2n}}, \quad \xi = \frac{\gamma_1}{\gamma_2},$$
$$\alpha_0 = 0{,}5\,(\mathring{\gamma}_2 + \mathring{\gamma}_3), \qquad \gamma_1 = -\mathring{\gamma}_2\mathring{\gamma}_3,$$
$$\gamma_2 = \max\,[\mathring{\gamma}_4(\Delta_2 - \Delta_1), \ \mathring{\gamma}_1(\Delta_2 - \Delta_1)] - \mathring{\gamma}_1\mathring{\gamma}_4,$$
$$\Delta_1 = \mathring{\gamma}_2 - \mathring{\gamma}_1, \qquad \Delta_2 = \mathring{\gamma}_4 - \mathring{\gamma}_3.$$

La méthode itérative (2) avec les paramètres indiqués sera appelée *méthode de Tchébychev.*

Voyons quelques cas particuliers. Soit $\Delta_1 = \Delta_2$, autrement dit les longueurs des segments $[\mathring{\gamma}_1, \mathring{\gamma}_2]$ et $[\mathring{\gamma}_3, \mathring{\gamma}_4]$ sont les mêmes. On a dans ce cas

$$\gamma_1 = -\mathring{\gamma}_2\mathring{\gamma}_3, \quad \gamma_2 = -\mathring{\gamma}_1\mathring{\gamma}_4, \quad \xi = \frac{\mathring{\gamma}_2\mathring{\gamma}_3}{\mathring{\gamma}_1\mathring{\gamma}_4}.$$

Montrons que dans le cas envisagé le jeu de paramètres τ_k construit est le meilleur. Cette assertion doit être démontrée vu que, lors de la construction des paramètres τ_k pour le schéma (2), on a posé n conditions (7), et, par suite, le choix des paramètres était soumis aux restrictions supplémentaires.

De (5) et de (8) on obtient que

$$x_{2n} = Q_{2n}(C)\, x_0 = P_n(\bar{C})\, x_0.$$

où

$$Q_{2n}(C) = \prod_{j=1}^{2n} (E - \tau_j C) = P_n(\bar{C}) = \prod_{j=1}^{n} (E - \omega_j \bar{C}). \qquad (24)$$

Considérons les polynômes algébriques correspondants $Q_{2n}(\mu)$ et $P_n(\lambda)$ ($\lambda = \mu^2 - 2\alpha\mu$). Si les paramètres ω_j sont choisis de la manière indiquée au théorème 1, on peut exprimer le polynôme $P_n(\lambda)$ de la façon suivante en fonction du polynôme de Tchébychev de 1ère espèce $T_n(\lambda)$ (voir ch. VI, § 2, point 1):

$$P_n(\lambda) = q_n T_n\left(\frac{1-\omega_0\lambda}{\rho_0}\right), \qquad P_n(0) = 1,$$

$$\max_{\gamma_1 \leqslant \lambda \leqslant \gamma_2} |P_n(\lambda)| = q_n.$$

Notons qu'aux points $\gamma_1 = \lambda_0 < \lambda_1 < \ldots < \lambda_n = \gamma_2$, où

$$\lambda_k = \frac{1 - \rho_0 \cos \frac{k\pi}{n}}{\omega_0}, \qquad k = 0,\ 1,\ \ldots,\ n,$$

le polynôme $P_n(\lambda)$ atteint des valeurs extrémales sur $[\gamma_1, \gamma_2]$:

$$P_n(\lambda_k) = (-1)^k q_n, \quad k = 0,\ 1,\ \ldots,\ n. \qquad (25)$$

Vu qu'en vertu de (24) on a l'égalité $Q_{2n}(\mu) = P_n(\lambda)$, où λ et μ sont liés par l'expression $\lambda = \mu^2 - 2\alpha\mu$, de (25) on tire

$$Q_{2n}(\mu_k^-) = Q_{2n}(\mu_k^+) = (-1)^k q_n, \; k = 0,\ 1,\ \ldots,\ n, \qquad (26)$$

où μ_k^- et μ_k^+ sont les racines de l'équation quadratique

$$\mu_k^2 - 2\alpha\mu_k - \lambda_k = 0, \quad k = 0,\ 1,\ \ldots,\ n. \qquad (27)$$

Ensuite, pour le cas considéré la transformation $\lambda = \lambda(\mu) = \mu^2 - 2\alpha\mu$ constitue une application de chacun de segments $[\mathring{\gamma}_1, \mathring{\gamma}_2]$ et $[\mathring{\gamma}_3, \mathring{\gamma}_4]$ sur le même segment $[\gamma_1, \gamma_2]$. Aux points $\mu = \mathring{\gamma}_2$, $\mu = \mathring{\gamma}_3$ correspond dans ce cas $\lambda = \gamma_1$, tandis qu'aux points $\mu = \mathring{\gamma}_1$ et $\mu = \mathring{\gamma}_4$ correspond $\lambda = \gamma_2$. Aussi les racines de l'équation (27) se disposent-elles de la façon suivante:

$$\mathring{\gamma}_1 = \mu_n^- < \mu_{n-1}^- < \ldots < \mu_0^- = \mathring{\gamma}_2, \;\; \mathring{\gamma}_3 = \mu_0^+ < \ldots < \mu_1^+ < \ldots < \mu_n^+ = \mathring{\gamma}_4.$$

Supposons maintenant que le jeu de paramètres τ_k construit dans le théorème 1 n'est pas le meilleur. Cela signifie qu'il existe un autre polynôme de degré non supérieur à $2n$ de la forme

$$\bar{Q}_{2n}(\mu) = \prod_{j=1}^{2n} (1 - \bar{\tau}_j \mu),$$

pour lequel

$$\max_{\mu \in \Omega} |\bar{Q}_{2n}(\mu)| < q_n, \qquad \Omega = [\mathring{\gamma}_1, \mathring{\gamma}_2] \cup [\mathring{\gamma}_3, \mathring{\gamma}_4].$$

Etudions la différence $R_{2n}(\mu) = Q_{2n}(\mu) - \bar{Q}_{2n}(\mu)$, qui est justement le polynôme de degré non supérieur à $2n$. En démontrant que le polynôme $R_{2n}(\mu)$ possède $2n + 2$ racines, on se convainc que l'hypothèse faite plus haut est erronée.

Pour esquisser cette démonstration, examinons les valeurs de $R_{2n}(\mu)$ aux points μ_k^-, $0 \leqslant k \leqslant n$. Vu que par hypothèse $-q_n < \bar{Q}_{2n}(\mu) < q_n$, $\mu \in \Omega$, on a

$$R_{2n}(\mu_k^-) = Q_{2n}(\mu_k^-) - \bar{Q}_{2n}(\mu_k^-) = (-1)^k q_n - \bar{Q}_{2n}(\mu_k^-)$$

et $R_{2n}(\mu_k^-) < 0$, si k est impair, et $R_{2n}(\mu_k^-) > 0$ si k est pair. Par conséquent, avec le passage de μ_k^- à μ_{k+1}^-, $0 \leqslant k \leqslant n-1$, le polynôme $R_{2n}(\mu)$ change de signe. Il existe donc sur le tronçon $[\mathring{\gamma}_1, \mathring{\gamma}_2]$ n racines de ce polynôme. De façon analogue, en étudiant les valeurs de $R_{2n}(\mu)$ aux points μ_k^+, $0 \leqslant k \leqslant n$, on démontre l'existence de n racines également sur le tronçon $[\mathring{\gamma}_3, \mathring{\gamma}_4]$. Ensuite, puisque

$$R_{2n}(\mathring{\gamma}_2) = R_{2n}(\mu_0^-) > 0, \quad R_{2n}(\mathring{\gamma}_3) = R_{2n}(\mu_0^+) > 0,$$
$$R_{2n}(0) = 0,$$

il existe dans l'intervalle $(\mathring{\gamma}_2, \mathring{\gamma}_3)$ soit deux racines différentes (l'une étant nulle) du polynôme $R_{2n}(\mu)$, soit zéro est une racine multiple. Donc, sur le segment $[\mathring{\gamma}_1, \mathring{\gamma}_4]$ le polynôme $R_{2n}(\mu)$ possède $2n + 2$ racines, ce qui est impossible.

Bref, pour le cas de $\Delta_1 = \Delta_2$ le jeu de paramètres τ_k construit au théorème 1 est le meilleur.

Soit maintenant $\Delta_1 \leqslant \Delta_2$. On a dans ce cas $\gamma_1 = -\mathring{\gamma}_2\mathring{\gamma}_3$, $\gamma_2 = \mathring{\gamma}_4(\Delta_2 - \Delta_1) - \mathring{\gamma}_1\mathring{\gamma}_4$. Comme $\gamma_2 = \mathring{\gamma}_4(\Delta_2 - \Delta_1) - \mathring{\gamma}_1\mathring{\gamma}_4 = \mathring{\gamma}_4(\mathring{\gamma}_4 - \mathring{\gamma}_3 - \mathring{\gamma}_2)$, γ_1 et γ_2 ne dépendent également pas de $\mathring{\gamma}_1$. Par conséquent, pour tout $\mathring{\gamma}_1$ de l'intervalle $\mathring{\gamma}_2 + \mathring{\gamma}_3 - \mathring{\gamma}_4 \leqslant \mathring{\gamma}_1 \leqslant \mathring{\gamma}_2$ on a le même jeu de paramètres τ_k, et la méthode itérative (2) converge avec une vitesse identique pour tout $\mathring{\gamma}_1$ de l'intervalle indiqué.

Notons en conclusion que le jeu de paramètres τ_k construit au théorème 1 sera également le meilleur au cas où $n = 1$, Δ_1 et Δ_2 n'étant pas forcément égaux. C'est le cas de la méthode itérative simple cyclique pour laquelle dans le schéma (2) $\tau_{2k-1} \equiv \tau_1$, $\tau_{2k} \equiv \tau_2$, $k = 1, 2, \ldots$, quant à τ_1 et τ_2, on les obtient à l'aide des formules du théorème 1 pour $n = 1$ $(\omega_1 = \omega_0)$

$$\tau_1 = -\alpha_0\omega_0 - \sqrt{\alpha_0^2\omega_0^2 + \omega_0}, \qquad \tau_2 = -\alpha_0\omega_0 + \sqrt{\alpha_0^2\omega_0^2 + \omega_0},$$

où $\omega_0 = 2/(\gamma_1 + \gamma_2)$. Vu que dans ce cas on a

$$x_{2n} = \prod_{j=1}^{n} (E - \omega_0 \bar{C})\, x_0 = (E - \omega_0 \bar{C})^n x_0,$$

$$\| E - \omega_0 \bar{C} \| \leqslant \rho_0, \qquad \rho_0 = \frac{1-\xi}{1+\xi}, \qquad \xi = \frac{\gamma_1}{\gamma_2},$$

on aboutira pour l'erreur z_{2n} du schéma itératif (2) à l'estimation

$$\| z_{2n} \|_D \leqslant \rho_0^n \| z_0 \|_D.$$

Etant donné qu'en vertu de (6) et (7) deux paramètres τ_1 et τ_2 sont remplacés par les paramètres ω_1 et α, ces derniers étant choisis de façon optimale ($\omega_1 = \omega_0$, $\alpha = \alpha_0$), les paramètres τ_1 et τ_2 s'avèrent effectivement les meilleurs pour la méthode itérative simple.

4. Méthodes itératives du type variationnel. On a étudié plus haut les méthodes itératives pour le cas de l'opérateur $DB^{-1}A$ autoadjoint dans la situation, où toutes les valeurs propres du problème (23) ne sont pas du même signe. Dans ce cas la convergence de la méthode itérative (2) était assurée par la construction d'un jeu spécial de paramètres d'itération. Examinons maintenant les méthodes itératives de l'aspect (2) dont la convergence avec le choix habituel des paramètres d'itération est assurée par la structure de l'opérateur B. Avec un tel procédé de construction des schémas itératifs on a déjà eu affaire dans la méthode de symétrisation de l'équation (voir ch. VI, § 4, point 4) ainsi que lors de l'étude des méthodes des moindres erreurs et des erreurs conjuguées au chapitre VIII.

Soit l'opérateur B de la forme

$$B = (A^*)^{-1}\tilde{B}, \tag{28}$$

où $\tilde{B}$ est un opérateur autoadjoint et défini positif arbitraire. En guise de l'opérateur D prenons $\tilde{B}$. Alors $DB^{-1}A = A^*A$, $C = \tilde{B}^{-1/2}A^*A\tilde{B}^{-1/2}$. Si l'opérateur B est non dégénéré et de signe indéterminé, alors l'opérateur C est tout de même défini positif. En outre, l'opérateur C est autoadjoint dans H. Aussi si γ_1 et γ_2 sont donnés dans les inégalités $\gamma_1 E \leqslant C \leqslant \gamma_2 E$, $\gamma_1 > 0$ ou dans les inégalités qui leur sont équivalentes

$$\gamma_1 \tilde{B} \leqslant A^*A \leqslant \gamma_2 \tilde{B}, \quad \gamma_1 > 0, \tag{29}$$

les paramètres τ_k de (2) peuvent être choisis suivant les formules de la méthode de Tchébychev à deux couches (voir ch. VI, § 2, point 1)

$$\tau_k = \frac{\tau_0}{1+\rho_0 \mu_k}, \quad \mu_k \in \mathfrak{M}_n^* = \left\{ -\cos \frac{(2i-1)\pi}{2n}, \quad 1 \leqslant i \leqslant n \right\},$$
$$k = 1, 2, \ldots, n, \tag{30}$$
$$\tau_0 = \frac{2}{\gamma_1+\gamma_2}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

On a ainsi le théorème 2.

Théorème 2. *Soit A l'opérateur non dégénéré. Pour la méthode itérative* (2), (28) *aux paramètres* (30), *où γ_1 et γ_2 sont donnés dans* (29), *on a l'estimation*

$$\| z_n \|_{\tilde{B}} \leqslant q_n \| z_0 \|_{\tilde{B}}, \quad q_n = \frac{2\rho_1^n}{1+\rho_1^{2n}}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}.$$

Si les constantes γ_1 et γ_2 de (29) sont soit inconnues, soit qu'elles peuvent être appréciées grossièrement, on peut recourir aux méthodes itératives du type variationnel étudiées au chapitre VIII.

Si pour le schéma (2), (28) les paramètres τ_k sont choisis suivant les formules

$$\tau_{k+1} = \frac{(r_k, r_k)}{(Aw_k, r_k)}, \quad k=0, 1, \ldots,$$

où $r_k = Ay_k - f$ est le résidu, tandis que w_k est la correction déduite de l'équation $\tilde{B}w_k = A^* r_k$, on aboutit à la méthode des moindres erreurs (voir ch. VIII, § 2, point 4). Pour l'erreur z_n, comme on le sait, on a l'estimation $\| z_n \|_{\tilde{B}} \leqslant \rho_0^n \| z_0 \|_{\tilde{B}}$, où ρ_0 est défini dans (30).

Si l'on s'adresse au schéma itératif à trois couches

$$By_{k+1} = \alpha_{k+1} (B - \tau_{k+1}A) y_k + (1 - \alpha_{k+1}) By_{k-1} + \tau_{k+1}\alpha_{k+1}f, \quad k \geqslant 1,$$

$$By_1 = (B - \tau_1 A) y_0 + \tau_1 f, \quad y_0 \in H,$$

où l'opérateur B est défini dans (28) et l'on choisit les paramètres d'itération α_{k+1} et τ_{k+1} suivant les formules

$$\tau_{k+1} = \frac{(r_k, r_k)}{(Aw_k, r_k)}, \quad k=0, 1, \ldots,$$

$$\alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(r_k, r_k)}{(r_{k-1}, r_{k-1})} \frac{1}{\alpha_k}\right)^{-1}, \quad k=1, 2, \ldots, \quad \alpha_1 = 1,$$

on aboutit à la méthode des erreurs conjuguées (voir ch. VIII, § 4, point 1). Pour l'erreur de cette méthode se vérifie l'estimation

$$\| z_n \|_{\tilde{B}} \leqslant q_n \| z_0 \|_{\tilde{B}}.$$

5. Exemples. Examinons l'application des méthodes construites plus haut à la recherche de la solution du problème discret de Dirichlet pour l'équation de Helmholtz dans un rectangle

$$\begin{aligned} &y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} + m^2 y = -f(x), \quad x \in \omega, \\ &y(x) = g(x), \quad x \in \gamma, \end{aligned} \tag{31}$$

où $\bar{\omega} = \{x_{ij} = (ih_1, jh_2),\ 0 \leqslant i \leqslant N_1,\ 0 \leqslant j \leqslant N_2,\ h_\alpha N_\alpha = l_\alpha,\ \alpha = 1, 2\}$, et γ la frontière du maillage $\bar{\omega}$.

Réduisons le problème (31) à l'équation opératorielle (1). H est dans le cas considéré l'espace des fonctions de mailles associées à ω avec produit scalaire

$$(u, v) = \sum_{x \in \omega} u(x) v(x) h_1 h_2, \quad u \in H, \quad v \in H.$$

Définissons l'opérateur R de la façon suivante: $Ry = -\Lambda\overset{\circ}{y}$, $y \in H$, $\overset{\circ}{y} \in \overset{\circ}{H}$ et $y(x) = \overset{\circ}{y}(x)$, $x \in \omega$, où $\overset{\circ}{H}$ est l'ensemble des fonctions de mailles données sur $\bar{\omega}$ et s'annulant sur γ, tandis que Λ est l'opérateur de différences de Laplace $\Lambda y = y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2}$. L'opérateur A s'obtient alors à partir de l'égalité $A = R - m^2E$. L'opérateur R étant autoadjoint dans H et possédant des valeurs propres

$$\lambda_k = \lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)}, \quad \lambda_{k_\alpha}^{(\alpha)} = \frac{4}{h_\alpha^2} \sin^2 \frac{k_\alpha \pi h_\alpha}{2l_\alpha}, \quad 1 \leqslant k_\alpha \leqslant N_\alpha - 1,$$

l'opérateur A est donc également autoadjoint dans H et ses valeurs propres μ_k s'expriment en fonction de λ_k suivant la formule

$$\mu_k = \lambda_k - m^2, \quad k = (k_1, k_2), \quad 1 \leqslant k_\alpha \leqslant N_\alpha - 1, \quad \alpha = 1, 2. \qquad (32)$$

Admettons que m^2 ne coïncide avec aucun λ_k. Désignons par λ_{m_1} et λ_{m_2} les valeurs de λ_k les plus rapprochées de m^2 respectivement par le bas et par le haut, c'est-à-dire

$$\lambda_{m_1} < m^2 < \lambda_{m_2}. \qquad (33)$$

Dans ce cas l'opérateur A est non dégénéré et de signe indéterminé.

Pour résoudre l'équation (1) munie de l'opérateur considéré A, examinons le schéma itératif explicite (2) ($B = E$). Si l'on pose $D = E$, l'opérateur $DB^{-1}A$ coïncide avec A et est autoadjoint dans H. Le choix des paramètres d'itération peut, dans ce cas, s'effectuer en recourant au théorème 1. De (23) il s'ensuit que l'information à priori nécessaire est fixée par les extrémités des segments $[\overset{\circ}{\gamma}_1, \overset{\circ}{\gamma}_2]$, $[\overset{\circ}{\gamma}_3, \overset{\circ}{\gamma}_4]$ des demi-axes positif et négatif sur lesquels se disposent les valeurs propres de l'opérateur A.

De (32) et (33) on tire $\overset{\circ}{\gamma}_1 = \delta - m^2$, $\overset{\circ}{\gamma}_2 = \lambda_{m_1} - m^2$, $\overset{\circ}{\gamma}_3 = \lambda_{m_2} - m^2$, $\overset{\circ}{\gamma}_4 = \Delta - m^2$, où

$$\delta = \min_k \lambda_k = \sum_{\alpha=1}^{2} \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2l_\alpha}, \quad \Delta = \max_k \lambda_k = \sum_{\alpha=1}^{2} \frac{4}{h_\alpha^2} \cos^2 \frac{\pi h_\alpha}{2l_\alpha}.$$

Cherchons maintenant γ_1, γ_2 et la quantité $\sqrt{\xi}$ qui définit le nombre d'itérations de la méthode étudiée, de sorte que $n \geqslant n_0(\varepsilon) = \ln(2/\varepsilon)/(2\sqrt{\xi})$. A partir des formules du théorème 1 on déduit

$$\gamma_1 = (m^2 - \lambda_{m_1})(\lambda_{m_2} - m^2),$$

$$\gamma_2 = \begin{cases} (\Delta - m^2)(\Delta + m^2 - \lambda_{m_1} - \lambda_{m_2}), & \lambda_{m_1} + \lambda_{m_2} \leqslant (\Delta + \delta), \\ (m^2 - \delta)(\lambda_{m_1} + \lambda_{m_2} - m^2 - \delta), & \lambda_{m_1} + \lambda_{m_2} \geqslant (\Delta + \delta). \end{cases}$$

Le rapport $\xi = \gamma_1/\gamma_2$ dépend de m^2. Pour se faire une idée sur la qualité de la méthode itérative considérée, cherchons la valeur de m^2 de l'intervalle $(\lambda_{m_1}, \lambda_{m_2})$ pour laquelle ξ est maximal. Il vient

$$m^2 = 0{,}5(\lambda_{m_1} + \lambda_{m_2}),$$

avec

$$\gamma_1 = \left(\frac{\lambda_{m_2} - \lambda_{m_1}}{2}\right)^2,$$

$$\gamma_2 = \begin{cases} (\Delta - m^2)^2, & 2m^2 \leqslant \Delta + \delta, \\ (m^2 - \delta)^2, & 2m^2 \geqslant \Delta + \delta. \end{cases}$$

Si m^2 est petit, c'est-à-dire si λ_{m_1} et λ_{m_2} sont proches de δ, alors $\gamma_1 = O(1)$ et $\gamma_2 = (\Delta - m^2)^2 = O\left(\frac{1}{|h|^4}\right)$. Dans ce cas $\xi = O(|h|^4)$ est le meilleur. Si λ_{m_1} et λ_{m_2} sont proches de Δ, on obtient de nouveau $\xi = O(|h|^4)$. C'est seulement pour le cas où λ_{m_1} et λ_{m_2} sont proches de $0{,}5\,(\Delta + \delta)$ qu'on obtient

$$\gamma_1 = O\left(\frac{1}{|h|^2}\right) \quad \text{et} \quad \gamma_2 = O\left(\frac{1}{|h|^4}\right),$$

de sorte que

$$\xi = O(|h|^2).$$

Notons que le schéma aux différences (31) peut être résolu au moyen des méthodes directes étudiées aux chapitres III, IV : soit par la méthode de réduction totale, soit par la méthode de séparation des variables. Les problèmes aux limites triponctuels obtenus dans ce cas doivent être résolus, à la différence du cas $m = 0$, par la méthode du balayage non monotone.

§ 2. Equations avec opérateur complexe

1. Méthode itérative simple. Soit donnée dans l'espace hilbertien complexe H l'équation

$$Au + qu = f, \tag{1}$$

où A est l'opérateur hermitien, tandis que $q = q_1 + iq_2$ est un nombre complexe. Pour une résolution approchée de l'équation (1), passons au schéma explicite à deux couches

$$\frac{y_{k+1} - y_k}{\tau} + (A + qE)\, y_k = f, \quad k = 0, 1, \ldots, \quad y_0 \in H, \tag{2}$$

où $\tau = \tau_1 + i\tau_2$ est un paramètre d'itération complexe.

On admettra que $q_1 \neq 0$, tandis que γ_1 et γ_2 sont les constantes des inégalités

$$\gamma_1 E \leqslant A \leqslant \gamma_2 E. \tag{3}$$

Etudions la convergence du schéma itératif (2) dans l'espace énergétique $H\ (D = E)$ et cherchons la valeur optimale du paramètre d'itération τ. En se servant de (1) et (2), écrivons l'équation de l'er-

reur $x_k = y_k - u$ sous la forme:

$$x_{k+1} = Sx_k, \quad k = 0, 1, \ldots, \qquad S = E - \tau C, \tag{4}$$

où

$$C = A + qE.$$

De (4) il vient

$$x_n = S^n x_0, \quad \| x_n \| \leqslant \| S^n \| \| x_0 \|. \tag{5}$$

Etudions l'opérateur de passage d'une itération à l'autre. Vu que l'opérateur A est hermitien, on a

$$C^* = A + \bar{q}E, \quad C^*C = CC^*,$$

autrement dit l'opérateur C est un opérateur normal. Donc l'opérateur S est également normal. Il est connu (voir ch. V, § 1, point 2) que pour l'opérateur normal S se vérifient les relations suivantes:

$$\| S^n \| = \| S \|^n, \quad \| S \| = \sup_{x \neq 0} \frac{|(Sx, x)|}{(x, x)}.$$

Aussi s'ensuit-il de (5) que le problème du choix du paramètre d'itération τ se réduit à la recherche de ce dernier sur la base de la condition du minimum de la norme de l'opérateur S.

Résolvons ce problème. De (3) il s'ensuivra que

$$z = \frac{(Cx, x)}{(x, x)} \in \Omega,$$

$$\Omega = \{z = z_1 + a(z_2 - z_1), \quad 0 \leqslant a \leqslant 1, \quad z_1 = \gamma_1 + q, \quad z_2 = \gamma_2 + q\},$$

où Ω est le tronçon dans un plan complexe, réunissant les points z_1 et z_2. Donc

$$\| S \| = \sup_{x \neq 0} \frac{|(Sx, x)|}{(x, x)} = \sup_{z \in \Omega} |1 - \tau z|$$

et le paramètre τ est recherché sur la base de la condition $\min\limits_{\tau} \max\limits_{z \in \Omega} |1 - \tau z|$.

Etudions la fonction $\varphi(z) = |1 - \tau z|$. Etant donné que les lignes du niveau $|1 - \tau z| = \rho_0$ sont des cercles concentriques de centre au point $1/\tau$ et de rayon $R = \rho_0/|\tau|$, pour acquérir la valeur optimale du paramètre $\tau = \tau_0$ les points z_1 et z_2 doivent se trouver sur la même ligne du niveau. Par conséquent, doivent se vérifier les égalités

$$|1 - \tau_0 z_1| = \rho_0, \quad |1 - \tau_0 z_2| = \rho_0,$$

avec $|1 - \tau_0 z| \leqslant \rho_0$ pour $z \in \Omega$.

Ecrivons ces égalités sous une forme équivalente

$$\left|\frac{1-\tau_0 z_2}{1-\tau_0 z_1}\right| = 1, \quad \rho_0 = \frac{|z_2 - z_1|}{|z_1|\left|\frac{z_2}{z_1} - \frac{1-\tau_0 z_2}{1-\tau_0 z_1}\right|}.$$

Comme, en vertu de la première égalité, avec la variation de τ_0 le nombre complexe

$$z = \frac{1-\tau_0 z_2}{1-\tau_0 z_1}$$

parcourt le cercle unitaire dans le plan complexe de centre à l'origine des coordonnées, ρ_0 sera minimal si l'égalité

$$\frac{1-\tau_0 z_2}{1-\tau_0 z_1} = -\frac{z_2}{z_1}\frac{|z_1|}{|z_2|}$$

est satisfaite. Cette condition fournit la valeur suivante de τ_0:

$$\tau_0 = \frac{|z_2|/z_2 + |z_1|/z_1}{|z_1| + |z_2|}. \tag{6}$$

Avec cette valeur $\tau = \tau_0$ on a pour la norme de l'opérateur S l'estimation

$$\| S \| = \rho_0 = \frac{|z_2 - z_1|}{|z_1| + |z_2|}, \tag{7}$$

en se servant de laquelle on obtient pour l'erreur x_n du schéma itératif (2) l'estimation

$$\| x_n \|_B < \rho_0^n \| x_0 \|_B. \tag{8}$$

Cherchons maintenant les conditions dont la satisfaction donne $\rho_0 < 1$. L'inégalité

$$|z_2 - z_1| = |z_1|\left|\frac{z_2}{|z_1|} - \frac{z_1}{|z_1|}\right| \leqslant |z_1|\left(1 + \frac{|z_2|}{|z_1|}\right) = |z_1| + |z_2|$$

étant satisfaite, l'égalité n'étant atteinte qu'avec la satisfaction de la condition

$$\frac{z_1}{|z_1|} = -\frac{z_2}{|z_1|}\frac{|z_1|}{|z_2|} = -\frac{z_2}{|z_2|}, \tag{9}$$

on a $\rho_0 < 1$ si (9) n'a pas lieu.

Dans le cas considéré $z_1 = \gamma_1 + q$ et $z_2 = \gamma_2 + q$. On tire sans peine de (9) que dans les deux cas $\rho_0 < 1$: soit $q_2 \neq 0$ et γ_1 et γ_2 sont quelconques, soit $q_2 = 0$, mais alors γ_1 et γ_2 sont soumis à la condition $(\gamma_1 + q_1)(\gamma_2 + q_1) > 0$. Posons que par la suite ces conditions sont remplies. Dans ce cas le processus d'itération (2) sera convergent.

Théorème 3. *Posons que A est un opérateur hermitien et que les conditions* (3) *sont remplies. Le processus itératif* (2) *avec le paramètre*

$$\tau = \tau_0 = \frac{1}{|\gamma_1 + q| + |\gamma_2 + q|}\left(\frac{|\gamma_1 + q|}{\gamma_1 + q} + \frac{|\gamma_2 + q|}{\gamma_2 + q}\right.$$

converge dans H, et pour l'erreur on a l'estimation (8), *où*

$$\rho_0 = \frac{|\gamma_2 - \gamma_1|}{|\gamma_1 + q| + |\gamma_2 + q|}.$$

R e m a r q u e. On a résolu plus haut le problème consistant dans la recherche du paramètre optimal τ sur la base de la condition $\min_\tau \max_{z \in \Omega} |1 - \tau z|$, où Ω est le tronçon du plan complexe réunissant deux points z_1 et z_2. On obtient sans peine la solution de ce problème également dans le cas où Ω est un cercle de centre au point z_0 et de rayon $r_0 < |z_0|$, c'est-à-dire ne comprenant pas l'origine des coordonnées. La solution du problème posé prend la forme

$$\tau_0 = \frac{1}{z_0}, \quad \sup_{z \in \Omega} |1 - \tau_0 z| = \rho_0 = \frac{r_0}{|z_0|} < 1.$$

Voyons maintenant comment s'applique la méthode construite à la recherche de la solution du problème de différences suivant:

$$\begin{gathered} \Lambda u - qu = -\varphi(x), \quad x \in \omega, \\ u(x) = g(x), \quad x \in \gamma, \quad q = q_1 + iq_2, \qquad (10) \\ \Lambda = \Lambda_1 + \Lambda_2, \quad \Lambda_\alpha u = (a_\alpha u_{\bar{x}_\alpha})_{x_\alpha}, \quad \alpha = 1, 2, \end{gathered}$$

où $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G}, 0 \leqslant i \leqslant N_1, 0 \leqslant j \leqslant N_2, h_\alpha N_\alpha = l_\alpha, \alpha = 1, 2\}$ est le maillage dans le rectangle $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha, \alpha = 1, 2\}$, quant aux coefficients $a_\alpha(x)$, ils sont réels et satisfont aux conditions

$$0 < c_1 \leqslant a_\alpha(x) \leqslant c_2, \quad x \in \bar{\omega}. \qquad (11)$$

Dans le cas considéré H est un espace des fonctions de mailles à valeurs complexes données sur ω avec produit scalaire

$$(u, v) = \sum_{x \in \omega} u(x) \bar{v}(x) h_1 h_2.$$

Le problème (10) s'écrira sous forme de l'équation (1), où l'opérateur A sera défini de façon habituelle: $Ay = -\Lambda \mathring{y}$, où $\mathring{y} \in \mathring{H}$, $y(x) = \mathring{y}(x)$ pour $x \in \omega$, $\mathring{y}(x) = 0$, $x \in \gamma$.

Pour résoudre l'équation (1) ainsi établie, prenons le schéma itératif explicite (2).

En faisant appel aux formules de différences de Green pour les fonctions à valeurs complexes, ainsi qu'aux inégalités (11), on se convainc que l'opérateur A est hermitien dans H, tandis que dans

les inégalités (3)

$$\gamma_1 = c_1 \sum_{\alpha=1}^{2} \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2l_\alpha},$$

$$\gamma_2 = c_2 \sum_{\alpha=1}^{2} \frac{4}{h_\alpha^2} \cos^2 \frac{\pi h_\alpha}{2l_\alpha}.$$

Si l'on choisit le paramètre d'itération τ en conformité avec le théorème 3, on aura alors pour l'erreur $x_n = y_n - u$ l'estimation (8), où ρ_0 est défini dans le théorème 3.

Dans le cas particulier, où $l_1 = l_2 = l$, $N_1 = N_2 = N$ et $q_1 = O(1)$, $q_2 = O(1)$, on obtient $\rho_0 = 1 - O(N^{-2})$. Donc pour atteindre la précision imposée ε il faut accomplir $n_0(\varepsilon) = O\left(N^2 \ln \frac{1}{\varepsilon}\right)$ itérations.

2. Méthode des directions alternées. Examinons de nouveau l'équation (1) et posons que l'opérateur A peut se représenter sous forme de somme de deux opérateurs hermitiens de permutation A_1 et A_2:

$$A = A_1 + A_2, \quad A_1A_2 = A_2A_1, \quad A_\alpha = A_\alpha^*, \quad \alpha = 1, 2. \quad (12)$$

Soient δ et Δ les bornes des opérateurs A_1 et A_2, c'est-à-dire

$$\delta E \leqslant A_\alpha \leqslant \Delta E, \quad \alpha = 1, 2. \quad (13)$$

Pour résoudre l'équation (1), adressons-nous au schéma itératif implicite à deux couches (2), où l'opérateur B est donné de la façon suivante:

$$B = (\omega E + A_1 + q_0E)(\omega E + A_2 + q_0E), \quad q_0 = 0{,}5q, \quad (14)$$

tandis que les paramètres τ et ω sont liés par la relation $\tau = 2\omega$. Le schéma itératif analogue a été obtenu au chapitre XI lors de la construction de la méthode des directions alternées. Notons que pour la recherche de y_{k+1} on peut recourir dans le schéma (2), (14) à l'algorithme suivant:

$$(\omega E + C_1)\, y_{k+1/2} = (\omega E - C_2) y_k + f,$$
$$(\omega E + C_2)\, y_{k+1} = (\omega - C_1)\, y_{k+1/2} + f,$$
$$k = 0, 1, \ldots,$$

où, pour abréger les notations, $C_\alpha = A_\alpha + q_0E$, $\alpha = 1, 2$.

Passons à l'étude de la convergence du schéma (2), (14) dans la norme H. Profitant de la permutabilité des opérateurs A_1 et A_2, on obtient l'équation pour l'erreur z_k

$$z_{k+1} = S_1S_2z_k, \quad k = 0, 1, \ldots, \quad (15)$$

$$S_\alpha = (\omega E + C_\alpha)^{-1}(\omega E - C_\alpha), \quad \alpha = 1, 2, \quad (16)$$

les opérateurs S_1 et S_2 étant permutables. De (15) il vient

$$z_n = S_1^n S_2^n z_0, \quad \| z_n \| \leqslant \| S_1^n \| \, \| S_2^n \| \, \| z_0 \|. \tag{17}$$

Apprécions la norme de l'opérateur S_α^n, $\alpha = 1, 2$. C_α étant un opérateur normal ($C_\alpha^* C_\alpha = C_\alpha C_\alpha^*$, $\alpha = 1, 2$), l'opérateur S_α est également normal. Donc $\| S_\alpha^n \| = \| S_\alpha \|^n$ et il suffit d'apprécier la norme de l'opérateur S_α même.

Comme la norme d'un opérateur normal est égale à son rayon spectral (voir ch. V, § 1, point 2), il s'ensuit de (16)

$$\| S_\alpha \| = \max_{\lambda_\alpha} \left| \frac{\omega - \lambda_\alpha}{\omega + \lambda_\alpha} \right|, \tag{18}$$

où λ_α sont les valeurs propres de l'opérateur C_α. En vertu des hypothèses (12) et (13) faites relativement aux opérateurs A_α, on obtient que $\lambda_\alpha \in \Omega = \{z = z_1 + a(z_2 - z_1), 0 \leqslant a \leqslant 1, z_1 = \delta + q_0, z_2 = \Delta + q_0\}$ pour $\alpha = 1, 2$. Donc, de (18) on tire que

$$\| S_\alpha \| \leqslant \max_{z \in \Omega} \left| \frac{\omega - z}{\omega + z} \right|, \quad \alpha = 1, 2. \tag{19}$$

Posons maintenant le problème de la recherche du paramètre ω sur la base de la condition du minimum du second membre de l'inégalité (19).

Considérons l'application linéaire fractionnaire

$$w = (\omega - z)/(\omega + z), \quad \omega \neq 0, \tag{20}$$

établissant la relation entre les points du plan z et les points du plan w. Il résulte des propriétés de la transformation (20) qu'aux cercles $| w | = \rho_0$ dans le plan w pour $\rho \neq 1$ correspondent des cercles du plan z, tandis qu'à un cercle unitaire correspond dans le plan z une droite passant par l'origine des coordonnées. Les points de la droite mentionnée possèdent un argument se différenciant de l'argument ω de $\pm\pi/2$.

Cherchons dans le plan z le centre et le rayon du cercle correspondant au cercle $| w | = \rho_0 \neq 1$ dans le plan w. Pour ce faire, exprimons z en fonction de w suivant (20):

$$z = \omega (1 - w)/(1 + w),$$

ensuite, en utilisant cette relation, calculons

$$\left| \frac{1 + | w |^2}{1 - | w |^2} \omega - z \right| = \left| \frac{1 + | w |^2}{1 - | w |^2} \omega - \frac{1 - w}{1 + w} \omega \right| = \\ = \frac{2 | \omega |}{| 1 - | w |^2 |} \cdot \frac{| w + | w |^2 |}{| 1 + w |}.$$

Comme

$$| w + | w |^2 | = | w + w\overline{w} | = | w | \, | -1 + \overline{w} | = | w | \, | 1 + w |,$$

on obtient finalement

$$\left|\frac{1+|w|^2}{1-|w|^2}\,\omega - z\right| = \frac{2\,|w|\,|\omega|}{|1-|w|^2|}.$$

Il résulte de ce qui précède qu'aux cercles $|w| = \rho_0 < 1$ correspondent les cercles du plan z de centre au point z_0 et de rayon R, où

$$z_0 = \frac{1+\rho_0^2}{1-\rho_0^2}\,\omega, \quad R = \frac{2\rho_0\,|\omega|}{1-\rho_0^2}. \tag{21}$$

Notons en outre qu'en vertu de l'univocité mutuelle de l'application (20) les égalités

$$\left|\frac{\omega-z}{\omega+z}\right| = \rho_0 < 1, \quad |z_0 - z| = R \tag{22}$$

sont équivalentes.

Revenons au problème posé. Etudions la fonction

$$\varphi(z) = |w| = \left|\frac{\omega-z}{\omega+z}\right|.$$

De ce qui vient d'être dit il s'ensuit que les lignes du niveau $\varphi(z) = \rho_0$ pour $\rho_0 < 1$ sont des cercles de centre au point z_0 et de rayon R, où z_0 et R sont définis dans (21). Pour des ρ_0 différents, ces cercles ne se coupent pas, le cercle correspondant à la plus petite valeur de ρ_0 se trouvant à l'intérieur du cercle correspondant à la plus grande valeur de ρ_0. On en déduit que pour la valeur optimale de $\omega = \omega_0$, les points z_1 et z_2 doivent se trouver sur une même ligne du niveau:

$$\left|\frac{\omega_0-z_1}{\omega_0+z_1}\right| = \rho_0 < 1, \quad \left|\frac{\omega_0-z_2}{\omega_0+z_2}\right| = \rho_0 < 1, \tag{23}$$

l'égalité

$$\max_{z\in\Omega}\left|\frac{\omega_0-z}{\omega_0+z}\right| = \rho_0$$

étant dans ce cas satisfaite. Le paramètre ω_0 doit être choisi sur la base de la condition du minimum de ρ_0.

Cherchons la valeur optimale de ω_0 et calculons ρ_0. De (23), en vertu de (22), il vient

$$|z_0 - z_1| = R_0, \quad |z_0 - z_2| = R_0,$$

$$z_0 = \frac{1+\rho_0^2}{1-\rho_0^2}\,\omega_0, \quad R_0 = \frac{2\rho_0\,|\omega_0|}{1-\rho_0^2}$$

ou

$$\left|\frac{z_0-z_2}{z_0-z_1}\right| = 1, \quad \frac{2\rho_0}{1+\rho_0^2} = \frac{R_0}{|z_0|} = \frac{|z_2-z_1|}{|z_1|\left|\frac{z_2}{z_1} - \frac{z_0-z_2}{z_0-z_1}\right|}. \tag{24}$$

Notons que ρ_0 est minimal quand $\frac{2\rho_0}{1+\rho_0^2}$ l'est. Or cela n'a lieu que si l'on exige que l'égalité

$$\frac{z_0-z_2}{z_0-z_1}=-\frac{z_2}{z_1}\frac{|z_1|}{|z_2|} \tag{25}$$

soit satisfaite. En portant cette expression dans (24), il vient

$$\frac{2\rho_0}{1+\rho_0^2}=\frac{|z_2-z_1|}{|z_1|+|z_2|}.$$

De là on obtient sans peine

$$\begin{aligned}\gamma_1 &= \frac{(1-\rho_0)^2}{1+\rho_0^2}=\frac{|z_1|+|z_2|-|z_2-z_1|}{|z_1|+|z_2|},\\ \gamma_2 &= \frac{(1+\rho_0)^2}{1+\rho_0^2}=\frac{|z_1|+|z_2|+|z_2-z_1|}{|z_1|+|z_2|},\quad \xi=\frac{\gamma_1}{\gamma_2}=\left(\frac{1-\rho_0}{1+\rho_0}\right)^2.\end{aligned} \tag{26}$$

Par conséquent,

$$\rho_0=\frac{1-\sqrt{\xi}}{1+\sqrt{\xi}},\qquad \frac{1-\rho_0^2}{1+\rho_0^2}=\sqrt{\gamma_1\gamma_2}$$

et, en outre,

$$z_0=\frac{1+\rho_0^2}{1-\rho_0^2}\omega_0=\frac{\omega_0}{\sqrt{\gamma_1\gamma_2}}.$$

En portant cette expression dans (25), on obtient la valeur optimale du paramètre ω_0:

$$\omega_0=\frac{|z_1|+|z_2|}{|z_1|/z_1+|z_2|/z_2}\sqrt{\gamma_1\gamma_2}. \tag{27}$$

Ainsi, on a obtenu pour des valeurs optimales de $\omega=\omega_0$ l'estimation de la norme de l'opérateur S_α: $\|S_\alpha\|\leqslant\rho_0$, $\alpha=1, 2$. En la portant dans (17), on obtient l'estimation pour l'erreur z_n:

$$\|z_n\|\leqslant\rho_0^{2n}\|z_0\|,\quad \rho_0=\frac{1-\sqrt{\xi}}{1+\sqrt{\xi}},\quad \xi=\frac{\gamma_1}{\gamma_2}. \tag{28}$$

En résonnant comme dans la méthode itérative simple, on aboutit aux inégalités $\gamma_1>0$ et $\rho_0<1$ qui auront lieu dans deux cas: soit pour $q_2\neq 0$, soit pour $q_2=0$, mais, toutefois, $(\delta+0{,}5q_1)\times\times(\Delta+0{,}5q_1)>0$.

On a ainsi démontré le théorème suivant.

Théorème 4. *Supposons les conditions* (12) *remplies*, δ *et* Δ *des inégalités* (13) *donnés, et soit* $q_2\neq 0$, *soit* $q_2=0$ *et* $(\delta+0{,}5q_1)\times\times(\Delta+0{,}5q_1)>0$. *Pour la méthode des directions alternées* (2), (14), *où le paramètre d'itération* $\omega=\omega_0$ *est choisi suivant la formule* (27), *tandis que* $\tau=2\omega_0$, *se vérifie l'estimation* (28), *où* γ_1 *et* γ_2 *sont définis dans* (26), *tandis que* $z_1=\delta+0{,}5q$ *et* $z_2=\Delta+0{,}5q$.

Remarque 1. La solution du problème $\min_{\omega} \max_{z \in \Omega} \left| \frac{\omega - z}{\omega + z} \right|$, où Ω est un cercle de centre au point z_0 et de rayon $r_0 < |z_0|$, prend la forme

$$\omega_0 = z_0 \sqrt{\gamma_1 \gamma_2}, \quad \rho_0 = \max_{z \in \Omega} \left| \frac{\omega - z}{\omega + z} \right| = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2},$$

où $\gamma_1 = 1 - r_0/|z_0|$, $\gamma_2 = 1 + r_0/|z_0|$.

Remarque 2. Si au lieu de l'inégalité (13) sont données les inégalités $\delta_\alpha E \leqslant A_\alpha \leqslant \Delta_\alpha E$, $\alpha = 1, 2$, il faut alors poser dans le théorème 4 $\delta = \min(\delta_1, \delta_2)$, $\Delta = \max(\Delta_1, \Delta_2)$.

§ 3. Méthodes itératives générales pour les équations avec opérateur dégénéré

1. Schémas itératifs au cas où l'opérateur B est non dégénéré. Supposons que dans l'espace hilbertien de dimension finie $H = H_N$ est donnée l'équation

$$Au = f \tag{1}$$

possédant un opérateur A linéaire dégénéré. Cette dernière condition signifie que l'égalité $Au = 0$ a lieu pour un certain $u \neq 0$. Rappelons (voir ch. V, § 2, point 2) l'information se rapportant à la résolution de l'équation (1).

Soit $\ker A$ le noyau de l'opérateur A, c'est-à-dire l'ensemble des éléments $u \in H$ pour lesquels $Au = 0$. Notons par $\operatorname{im} A$, image de l'opérateur A, l'ensemble des éléments de la forme $y = Au$, où $u \in H$. On sait qu'il y a lieu des développements orthogonaux suivants de l'espace H en sommes directes de deux sous-espaces:

$$H = \ker A \oplus \operatorname{im} A^*, \quad H = \ker A^* \oplus \operatorname{im} A. \tag{2}$$

Cela signifie que tout élément $u \in H$ peut se représenter sous forme de $u = \bar{u} + \tilde{u}$, où $\bar{u} \in \operatorname{im} A^*$ et $\tilde{u} \in \ker A$ avec $(\bar{u}, \tilde{u}) = 0$. De façon analogue, $u = \bar{u} + \tilde{u}$, où $\bar{u} \in \operatorname{im} A$ et $\tilde{u} \in \ker A^*$, $(\bar{u}, \tilde{u}) = 0$.

Soit dans l'équation (1) $f = \bar{f} + \tilde{f}$, où $\bar{f} \in \operatorname{im} A$, $\tilde{f} \in \ker A^*$. On appelle solution généralisée (1) l'élément $u \in H$ pour lequel $Au = \bar{f}$; elle garantit un minimum à la fonctionnelle $\| Au - f \|$. La solution généralisée n'est pas unique et se détermine à la précision de l'élément $\ker A$ près. La solution normale est une solution généralisée possédant une norme minimale. La solution normale est unique et appartient à $\operatorname{im} A^*$.

Notre objectif est de construire les méthodes permettant de rechercher de façon approchée la solution normale de l'équation (1). On exigera dans ce cas que la solution approchée, de même que la solution normale précise, appartienne à l'espace $\operatorname{im} A^*$.

Pour résoudre le problème posé, utilisons le schéma implicite à deux couches

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}}+Ay_k=f,\quad k=0,\ 1,\ \ldots,\quad y_0\in H. \tag{3}$$

Etudions d'abord le cas de l'opérateur B non dégénéré dans H. Les exigences générales envers le processus itératif sont:

a) les itérations sont effectuées suivant le schéma (3), l'approximation $y_n\in \operatorname{im} A^*$, quant aux approximations intermédiaires y_k, elles peuvent appartenir à H;

b) la structure concrète des sous-espaces $\ker A$, $\ker A^*$, $\operatorname{im} A$ et $\operatorname{im} A^*$ n'est pas utilisée au cours des itérations.

Cherchons les conditions imposées à l'opérateur B, l'approximation initiale y_0 et les paramètres τ_k, $k=1,\ 2,\ \ldots,\ n$ qui garantissent la satisfaction des exigences formulées plus haut.

C o n d i t i o n s 1. Soit l'opérateur B tel que

$$Bu\in\ker A^*,\quad \text{si}\quad u\in\ker A, \tag{4}$$

$$Bu\in\operatorname{im} A,\quad \text{si}\quad u\in\operatorname{im} A^*. \tag{5}$$

On a le lemme 2.

L e m m e 2. *Si pour les opérateurs A et B se justifient les égalités*

$$A^*B=CA,\qquad BA^*=AD, \tag{6}$$

où C et D sont des opérateurs quelconques, les conditions (4) *et* (5) *peuvent être considérées comme satisfaites.*

En effet, supposons les égalités (6) satisfaites. Si $u\in\ker A$, alors $Au=0$ et, par suite, $A^*Bu=CAu=0$. Pour cette raison $Bu\in\ker A^*$, et la condition (4) est remplie. Supposons à présent que $u\in\operatorname{im} A^*$, c'est-à-dire que $u=A^*v$, où $v\in H$. Alors on a $Bu=BA^*v=ADv\in\operatorname{im} A$. Par conséquent, la condition (5) est remplie. Le lemme est démontré.

C o r o l l a i r e. *Pour le cas de $A=A^*$ les conditions du lemme* 2 *seront satisfaites si les opérateurs A et B sont permutables*: $AB=BA$.

Fournissons encore une série de propositions découlant de (4) et de (5).

L e m m e 3. *Soient les conditions* (4) *et* (5) *remplies. Alors*

$$B^{-1}u\in\ker A,\quad \textit{si}\quad u\in\ker A^*, \tag{7}$$

$$B^{-1}u\in\operatorname{im} A^*,\quad \textit{si}\quad u\in\operatorname{im} A, \tag{8}$$

et l'opérateur AB^{-1} n'est pas dégénéré sur $\operatorname{im} A$.

De fait, soit $u\in\ker A^*$ et $u\neq 0$. Posons $v=B^{-1}u$ et admettons que $v\in\operatorname{im} A^*$. Alors en vertu de (5) $u=Bv\in\operatorname{im} A$. Mais comme $u\neq 0$ et les espaces $\operatorname{im} A$ et $\ker A^*$ sont orthogonaux, l'hypothèse

faite est erronée. Donc $v = B^{-1}u \in \ker A$ et (7) est démontré. De façon analogue on démontre (8).

Montrons maintenant que AB^{-1} n'est pas dégénéré sur le sous-espace $\operatorname{im} A$. En effet, soit $u \in \operatorname{im} A$. Alors en vertu de (8) $B^{-1}u \in \operatorname{im} A^*$ et, par suite, $B^{-1}u \perp \ker A$. De là on obtient que $AB^{-1}u \neq 0$, et donc $(AB^{-1}u, AB^{-1}u) > 0$. Le lemme est démontré.

Revenons maintenant au schéma (3) et voyons ce que donne la condition 1. En conformité avec le développement de H en forme de (2), représentons f et y_k pour tout k sous l'aspect

$$\begin{aligned} f &= \bar{f} + \tilde{f}, & \bar{f} &\in \operatorname{im} A, & \tilde{f} &\in \ker A^*, \\ y_k &= \bar{y}_k + \tilde{y}_k, & \bar{y}_k &\in \operatorname{im} A^*, & \tilde{y}_k &\in \ker A. \end{aligned} \tag{9}$$

En se servant de (9), écrivons le schéma (3) de la façon suivante:

$$B\frac{\bar{y}_{k+1} - \bar{y}_k}{\tau_{k+1}} + B\frac{\tilde{y}_{k+1} - \tilde{y}_k}{\tau_{k+1}} + A\bar{y}_k = \bar{f} + \tilde{f}, \quad k = 0, 1, \ldots \tag{10}$$

A partir de (4) et (5) on obtient que le premier terme du premier membre de (10) appartient à $\operatorname{im} A$, tandis que le second à $\ker A^*$. De (10) on tire donc l'équation

$$B\frac{\bar{y}_{k+1} - \bar{y}_k}{\tau_{k+1}} + A\bar{y}_k = \bar{f}, \quad k = 0, 1, \ldots, \quad \bar{y}_0 \in \operatorname{im} A^* \tag{11}$$

pour la composante $\bar{y}_k \in \operatorname{im} A^*$ et l'équation

$$B\frac{\tilde{y}_{k+1} - \tilde{y}_k}{\tau_{k+1}} = \tilde{f}, \quad k = 0, 1, \ldots, \quad \tilde{y}_0 \in \ker A \tag{12}$$

pour la composante $\tilde{y}_k \in \ker A$.

Cherchons les conditions qui, une fois satisfaites, donnent $y_n \in \operatorname{im} A^*$. De (9) il résulte que si $\tilde{y}_n = 0$, $y_n = \bar{y}_n \in \operatorname{im} A^*$. De (12) tirons l'expression explicite de $\tilde{y}_n$ et égalons-la à zéro. Alors l'exigence formulée dans a) sera satisfaite.

De (12) il vient

$$\tilde{y}_{k+1} = \tilde{y}_k + \tau_{k+1}B^{-1}\tilde{f} = \ldots = \tilde{y}_0 + \sum_{j=1}^{k+1} \tau_j B^{-1}\tilde{f}.$$

D'où suivent les conditions 2.

C o n d i t i o n s 2. Soit $y_0 = A^*\varphi$, où $\varphi \in H$, quant aux paramètres τ_k, $k = 1, 2, \ldots, n$, ils satisfont à l'exigence

$$\sum_{j=1}^{n} \tau_j = 0, \tag{13}$$

si $f \in H$. Si $f \perp \ker A^*$, la limitation ne joue pas pour les paramètres τ_k.

Eclairons le choix de l'approximation initiale y_0. Vu que pour tout $\varphi \in H$ on a $y_0 = A^*\varphi \in \text{im } A^*$, dans le développement (9) on a $\tilde{y}_0 = 0$ et $\bar{y}_0 = y_0$. En particulier, en choisissant $\varphi = 0$, on obtient l'approximation initiale $y_0 = 0$.

Ainsi, si les conditions 2 sont remplies, alors $y_n = \bar{y}_n$. Le processus itératif (3) convergera donc et fournira la solution normale approchée de l'équation (1) au cas où le processus d'itération (11) converge, c'est-à-dire si la suite $\bar{y}_k$ converge vers la solution normale $\bar{u}$.

R e m a r q u e 1. Les conditions 2 permettent de dégager de l'approximation itérative y_n sa projection sur im A^*, c'est-à-dire de trouver $\bar{y}_n$ sans faire appel aux sous-espaces ker A, im A, ker A^* ou im A^*. Ensuite, si l'on sait que $\sum\limits_{j=1}^{n} \tau_j \| B^{-1}\tilde{f} \|$ est petit, c'est-à-dire que $\| \tilde{y}_n \|$ est petit, alors, en vertu de l'égalité $\| y_n - \bar{u} \| = \| \bar{y}_n - \bar{u} \| + \| \tilde{y}_n \|$, on peut prendre en qualité de solution approchée y_n et s'abstenir de la limitation (13). Dans ce cas $y_n \in$ $\in$ im A^*.

R e m a r q u e 2. A la condition que tous les éléments du sous-espace ker A^* sont connus, on peut se borner à n'étudier que le cas de $f \perp \ker A^*$, en soustrayant, si nécessaire, de f sa projection sur ker A^*. Si l'on considère le processus itératif non stationnaire

$$B_{k+1}\frac{y_{k+1}-y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0,\ 1,\ \ldots,\ y_0 = A^*\varphi,$$

et si l'on exige que les conditions 1 soient remplies, où B est remplacé par B_k, $k = 1, 2, \ldots$, alors tous les $y_k \in \text{im } A^*$ et il n'est plus nécessaire d'imposer aucune autre limitation à τ_k.

2. Méthode itérative des moindres résidus. Examinons maintenant le problème du choix des paramètres d'itération τ_k pour le schéma (3). On posera que l'opérateur B satisfait aux conditions 1, tandis que les limitations imposées au choix de l'approximation initiale y_0 et aux paramètres τ_k sont définies par les conditions 2.

On a montré plus haut que les paramètres τ_k doivent être choisis sur la base de la condition de convergence du processus itératif (11) vers la solution normale $\bar{u}$ de l'équation (1).

Etudions le schéma itératif (11). Notons d'abord que l'opérateur $D = A^*A$ est défini positif sur im A^*. En effet, soit $u \in \text{im } A^*$ et $u \neq 0$. Comme $u \perp \ker A$, alors $Au \neq 0$ et, par suite, $(Du, u) = \| Au \|^2 > 0$. L'opérateur D engendre l'espace énergétique H_D composé d'éléments im A^*, où le produit scalaire se détermine de la façon habituelle: $(u, v)_D = (Du, v)$, $u \in \text{im } A^*$, $v \in \text{im } A^*$.

Abordons maintenant le problème du choix des paramètres τ_{k+1} dans le schéma (11) sur la base de la condition du minimum de

$\| \bar{z}_{k+1} \|_D$, ou $\bar{z}_{k+1}$ est l'erreur: $\bar{z}_{k+1} = \bar{y}_{k+1} - \bar{u}$, $A\bar{u} = \bar{f}$ et $\bar{u}$ étant une solution normale de l'équation (1).

On obtient pour l'erreur $\bar{z}_k \in \operatorname{im} A^*$ à partir de (11) l'équation suivante

$$\bar{z}_{k+1} = (E - \tau_{k+1} B^{-1} A) \bar{z}_k. \tag{14}$$

De là on tire

$$\| \bar{z}_{k+1} \|_D^2 = \| \bar{z}_k \|_D^2 - 2\tau_{k+1} (AB^{-1}A\bar{z}_k, A\bar{z}_k) + \tau_{k+1}^2 \| AB^{-1}A\bar{z}_k \|^2.$$

Notons qu'en vertu du lemme 3 $\| AB^{-1}A\bar{z}_k \| > 0$, $(A\bar{z}_k \in \operatorname{im} A)$. Le minimum $\| \bar{z}_{k+1} \|_D^2$ est donc atteint pour

$$\tau_{k+1} = \frac{(AB^{-1}A\bar{z}_k, A\bar{z}_k)}{(AB^{-1}A\bar{z}_k, AB^{-1}A\bar{z}_k)} \tag{15}$$

et est égal à

$$\| \bar{z}_{k+1} \|_D^2 = \rho_{k+1}^2 \| \bar{z}_k \|_D^2, \quad \rho_{k+1}^2 = 1 - \frac{(AB^{-1}A\bar{z}_k, A\bar{z}_k)^2}{\| AB^{-1}A\bar{z}_k \|^2 \| A\bar{z}_k \|^2}. \tag{16}$$

La formule (15) ne convient pas encore pour les calculs, car elle renferme des grandeurs inconnues. Transformons-la. En utilisant le développement (9), il vient

$$A\bar{z}_k = A\bar{y}_k - \bar{f} = Ay_k - \bar{f} = r_k + \tilde{f}, \tag{17}$$

où $r_k = Ay_k - f$ est le résidu. Vu que $\tilde{f} \in \ker A^*$, alors, en vertu du lemme 3, $B^{-1}\tilde{f} \in \ker A$ et, par conséquent, $AB^{-1}A\bar{z}_k = AB^{-1}r_k$. En portant cette expression, de même que (17), dans (15) et compte tenu de l'égalité $A^*\tilde{f} = 0$, il vient

$$\tau_{k+1} = \frac{(AB^{-1}r_k, r_k)}{(AB^{-1}r_k, AB^{-1}r_k)} = \frac{(Aw_k, r_k)}{(Aw_k, Aw_k)}, \tag{18}$$

où la correction w_k s'obtient à partir de l'équation $Bw_k = r_k$.

Notons que (18) coïncide avec la formule du paramètre d'itération τ_{k+1} de la méthode des moindres résidus étudiée au chapitre VIII pour l'équation avec paramètre A non dégénéré.

Apprécions maintenant la vitesse de convergence de la méthode construite. Multiplions (14) à gauche par A, calculons la norme des parties gauche et droite et, compte tenu de ce que $\| A\bar{z}_k \| = \| \bar{z}_k \|_D$, on aboutit à l'estimation suivante:

$$\| \bar{z}_{k+1} \|_D \leqslant \| E - \tau_{k+1} AB^{-1} \|_{\operatorname{im} A} \| \bar{z}_k \|_D \tag{19}$$

pour tout τ_{k+1}. De (16) et (19) on obtient pour tout τ_{k+1}

$$\rho_{k+1} \leqslant \| E - \tau_{k+1} AB^{-1} \|_{\operatorname{im} A}. \tag{20}$$

Si l'on note

$$\rho_0 = \min_{\tau} \| E - \tau AB^{-1} \|_{\operatorname{im} A},$$

il s'ensuivra de (16) et (20) une estimation pour l'erreur

$$\| \bar{z}_{k+1} \|_D \leqslant \rho_0 \| \bar{z}_k \|_D. \tag{21}$$

La notation $\| S \|_{\text{im} A}$ est utilisée ici pour désigner la norme de l'opérateur S dans le sous-espace im A.

Si $\rho_0 < 1$, la méthode itérative (11), (18) convergera dans H_D, et à partir de (21) on obtiendra que

$$\| \bar{z}_k \|_D \leqslant \rho_0^k \| \bar{z}_0 \|_D, \quad k = 0, 1, \ldots \tag{22}$$

Il ne reste qu'à subordonner le choix des paramètres τ_k à la condition (13), si $\tilde{f} \neq 0$. Procédons pour cela de la façon suivante. Effectuons, suivant le schéma (3), $(n - 1)$-ième itération en choisissant $y_0 = A^*\varphi$, où $\varphi \in H$, et en se servant pour les paramètres τ_{k+1}, $k = 0, 1, \ldots, n - 2$, de la formule (18). Effectuons encore une itération en choisissant

$$\tau_n = - \sum_{j=1}^{n-1} \tau_j.$$

La condition (13) est alors satisfaite et, par conséquent, $y_n = \bar{y}_n$. Apprécions maintenant la norme d'erreur $z_n = y_n - \bar{u}$ dans H_D. Comme $y_n = \bar{y}_n$, de (11) il vient

$$y_n = \bar{y}_{n-1} - \tau_n B^{-1} (A\bar{y}_{n-1} - \bar{f}) = \bar{y}_{n-1} - \tau_n B^{-1} A \bar{z}_{n-1}.$$

De là

$$z_n = \bar{z}_{n-1} - \tau_n B^{-1} A \bar{z}_{n-1}$$

et, après multiplication par A, il vient

$$Az_n = (E - \tau_n AB^{-1}) A\bar{z}_{n-1}.$$

En calculant la norme, on obtient l'estimation

$$\| z_n \|_D \leqslant \| E - \tau_n AB^{-1} \|_{\text{im} A} \| \bar{z}_{n-1} \|_D.$$

En y portant (22) et compte tenu de ce qu'en vertu du choix de y_0 on a l'égalité $\bar{y}_0 = y_0$, on obtient

$$\| y_n - \bar{u} \|_D \leqslant \| E - \tau_n AB^{-1} \|_{\text{im} A} \rho_0^{n-1} \| y_0 - \bar{u} \|_D. \tag{23}$$

Examinons des cas particuliers.

1) Soit $B = E$, l'opérateur A étant autoadjoint dans H. γ_1 et γ_2 sont les constantes des inégalités

$$\gamma_1 (x, x) \leqslant (Ax, x) \leqslant \gamma_2 (x, x), \quad \gamma_1 > 0, \quad Ax \neq 0. \tag{24}$$

Dans ce cas les conditions 1 sont remplies.

Cherchons ρ_0 et apprécions la norme de l'opérateur dans (23). L'opérateur A étant autoadjoint dans H, en utilisant (24), on aboutit à

$$\| E - \tau A \|_{\text{im} A} = \sup_{Au \neq 0} \left| 1 - \tau \frac{(Au, u)}{(u, u)} \right| \leqslant \max_{\gamma_1 \leqslant t \leqslant \gamma_2} | 1 - \tau t |.$$

On s'est déjà heurté au calcul du maximum mentionné et au choix de τ sur la base de son minimum au chapitre VI lors de l'étude de la méthode itérative simple. On a alors trouvé que

$$\min_{\tau} \max_{\gamma_1 \leqslant t \leqslant \gamma_2} |1-\tau t| = \rho_0 = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

Ainsi, ρ_0 est obtenu. Ensuite, avec $B = E$, la formule (15) du paramètre τ_{k+1} s'écrit sous la forme

$$\tau_{k+1} = \frac{(Ax, x)}{(Ax, Ax)}, \quad x = A\bar{z}_k \in \operatorname{im} A.$$

Vu que $A = A^*$ et $\gamma_1 > 0$, les inégalités (24) sont équivalentes aux inégalités suivantes (voir ch. V, § 1, point 3):

$$\gamma_1 (Ax, x) \leqslant (Ax, Ax) \leqslant \gamma_2 (Ax, x), \quad Ax \neq 0.$$

Pour cette raison les paramètres τ_k pour $k \leqslant n-1$ vérifient les inégalités $1/\gamma_2 \leqslant \tau_k \leqslant 1/\gamma_1$. De là on tire l'estimation

$$0 < \tau_n = \sum_{j=1}^{n-1} \tau_j \leqslant \frac{n-1}{\gamma_1}. \tag{25}$$

Apprécions la norme de l'opérateur dans (23). Compte tenu de (24) et (25), on obtient

$$\| E - \tau_n A \|_{\operatorname{im} A} \leqslant \max_{\gamma_1 \leqslant t \leqslant \gamma_2} |1 - \tau_n t| =$$

$$= 1 - \tau_n \gamma_2 \leqslant 1 + (n-1) \frac{\gamma_2}{\gamma_1} = 1 + (n-1) \frac{1+\rho_0}{1-\rho_0}.$$

Portons cette estimation dans (23) et l'on trouve

$$\| y_n - \bar{u} \|_D \leqslant \rho_0^{n-1} \left[1 + (n-1) \frac{1+\rho_0}{1-\rho_0} \right] \| y_0 - \bar{u} \|_D. \tag{26}$$

2) Soient $B = B^*$, $A = A^*$ et $AB = BA$. γ_1 et γ_2 sont les constantes des inégalités

$$\gamma_1 (Bx, x) \leqslant (Ax, x) \leqslant \gamma_2 (Bx, x), \quad \gamma_1 > 0, \quad Ax \neq 0. \tag{27}$$

Dans ce cas les conditions 1 sont satisfaites, l'opérateur AB^{-1} est autoadjoint dans H et l'on peut montrer que pour l'erreur de la méthode (3), (18) se justifie l'estimation (26).

3) Posons que les opérateurs B^*A et AB^* sont autoadjoints dans H, γ_1 et γ_2 étant les constantes de (27). Dans ce cas, en vertu du lemme 2, les conditions 1 sont satisfaites. En outre, l'opérateur AB^{-1} sera autoadjoint dans H. On peut montrer que dans ce cas l'estimation (26) joue également.

3. Méthode à paramètres de Tchébychev. Voyons maintenant les méthodes itératives (3) dont les paramètres τ_k sont choisis en se servant de l'information à priori sur les opérateurs A et B.

Introduisons d'abord certaines propositions auxiliaires qui nous seront nécessaires dans l'exposé ultérieur.

L e m m e 4. *Soient satisfaites les conditions*

$$A = A^* \geqslant 0, \quad B = B^* > 0, \quad AB = BA \tag{28}$$

ainsi que données les constantes γ_1 *et* γ_2 *dans les inégalités*

$$\gamma_1 (Bx, x) \leqslant (Ax, x) \leqslant \gamma_2 (Bx, x), \quad \gamma_1 > 0, \quad Ax \neq 0. \tag{29}$$

Notons par D *l'un des opérateurs* A, B *ou* $AB^{-1}A$ *et définissons sur le sous-espace* im A *l'opérateur* C

$$C = D^{-1/2} (DB^{-1}A) D^{-1/2}.$$

L'opérateur C *est autoadjoint dans* im A *et vérifie les inégalités*

$$0 < \gamma_1 (x, x) \leqslant (Cx, x) \leqslant \gamma_2 (x, x), \quad x \in \text{im } A. \tag{30}$$

En effet, de (28) et du corollaire du lemme 2 on déduit que les conditions 1 sont satisfaites. Ensuite, l'opérateur D est autoadjoint dans H et défini positif sur im A. A titre d'exemple démontrons que l'opérateur $D = AB^{-1}A$ est défini positif. Posons $u \in$ im A et $u \neq 0$. Comme $(Du, u) = (B^{-1}Au, Au)$ et l'opérateur B^{-1} est défini positif en vertu du fait que l'opérateur B est borné et défini positif, $(Du, u) \geqslant 0$, et ne peut s'annuler que si la condition $Au = 0$ est satisfaite. Or cela contredit les hypothèses faites.

L'opérateur D étant une application de im A sur im A, il existe alors un $D^{-1/2}$ qui est également une application de cet espace sur lui-même. On peut donc définir sur im A l'opérateur C mentionné dans le lemme. Le passage de (29) à (30) se démontre également de la même façon que cela a été réalisé au chapitre VI, § 2, point 3. Le lemme est démontré.

L e m m e 5. *Supposons que soient remplies les conditions*

$$B^*A = A^*B, \quad AB^* = BA^* \tag{31}$$

et données γ_1 *et* γ_2 *dans* (29). *Posons* $C_1 = AB^{-1}$ *et* $C_2 = B^{-1}A$. *Les opérateurs* C_1 *et* C_2 *sont autoadjoints dans* H *et vérifient les inégalités*

$$\gamma_1 (x, x) \leqslant (C_1x, x) \leqslant \gamma_2 (x, x), \ \gamma_1 > 0, \ x \in \text{im } A, \tag{32}$$

$$\gamma_1 (x, x) \leqslant (C_2x, x) \leqslant \gamma_2 (x, x), \ \gamma_1 > 0, \ x \in \text{im } A^*. \tag{33}$$

Le fait que les opérateurs C_1 et C_2 sont autoadjoints s'ensuit de (31). Montrons-le pour l'exemple (32). Examinons le problème sur les valeurs propres

$$AB^{-1}v - \lambda v = 0, \quad v \in H. \tag{34}$$

L'opérateur AB^{-1} étant autoadjoint dans H, il existe alors un système orthonormé des fonctions propres du problème (34) $\{v_1, v_2, \ldots$

$\ldots, v_p, v_{p+1}, \ldots, v_N\}$. Soient $v_1, \ldots, v_p$ les fonctions correspondant à la valeur propre $\lambda = 0$, et $v_{p+1}, \ldots, v_N$ correspondant aux valeurs non nulles de λ. On voit sans peine que $v_i \in \ker A^*$, $1 \leqslant i \leqslant p$, $v_i \in \operatorname{im} A$, $p+1 \leqslant i \leqslant N$ et, en vertu du développement de H en sous-espaces (2), les fonctions $v_{p+1}, \ldots, v_N$ constituent une base dans $\operatorname{im} A$. On a alors pour $x \in \operatorname{im} A$

$$x = \sum_{k=p+1}^{N} a_k v_k, \quad C_1 x = \sum_{k=p+1}^{N} \lambda_k a_k v_k,$$

et en vertu de l'orthogonalité des fonctions propres

$$(x, x) = \sum_{k=p+1}^{N} a_k^2, \quad (C_1 x, x) = \sum_{k=p+1}^{N} \lambda_k a_k^2.$$

On obtient de là les inégalités

$$\min_{p+1 \leqslant k \leqslant N} \lambda_k (x, x) \leqslant (C_1 x, x) \leqslant \max_{p+1 \leqslant k \leqslant N} \lambda_k (x, x).$$

Il reste à trouver les valeurs propres minimale et maximale correspondant aux fonctions propres du problème (34), appartenant à $\operatorname{im} A$. Ecrivons (34) sous la forme

$$Au_k - \lambda_k B u_k = 0, \quad p+1 \leqslant k \leqslant N, \tag{35}$$

où $u_k = B^{-1} v_k \in \operatorname{im} A^*$ et, partant, $Au_k \neq 0$. En multipliant scalairement (35) par u_k et utilisant (29), on obtient que

$$\min_{p+1 \leqslant k \leqslant N} \lambda_k = \gamma_1, \quad \max_{p+1 \leqslant k \leqslant N} \lambda_k = \gamma_2.$$

Les inégalités (32) sont démontrées. La justesse de (33) s'établit de la même façon. Le lemme est démontré.

Revenons maintenant au problème du choix des paramètres d'itération pour le schéma (3). Compte tenu des conditions 2, écrivons ce dernier sous la forme suivante:

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad y_0 \in A^* \varphi, \quad \sum_{j=1}^{n} \tau_j = 0. \tag{36}$$

Si les conditions 1 sont satisfaites, les paramètres τ_k doivent être choisis sur la base de la condition de convergence du schéma (11) avec la restriction susmentionnée concernant la somme τ_j.

Examinons l'équation (14) de l'erreur du schéma (11). Si les conditions du lemme 4 sont remplies, alors, en posant dans (14) $z_k = D^{-1/2} x_k$, où D est l'un des opérateurs du lemme 4, on obtient l'équation suivante pour l'erreur équivalente:

$$x_{k+1} = (E - \tau_{k+1} C) x_k, \; k = 0, 1, \ldots, \; x_k \in \operatorname{im} A. \tag{37}$$

L'opérateur C est également défini dans le lemme 4.

Si les conditions du lemme 5 sont remplies, alors, en posant $B\bar{z}_k = x_k$ ou $A\bar{z}_k = x_k$, on obtient l'équation

$$x_{k+1} = (E - \tau_{k+1}C_1)\, x_k,\ k = 0,\ 1,\ \ldots,\ x_k \in \operatorname{im} A. \tag{38}$$

Dans ce cas $\| x_k \| = \| \bar{z}_k \|_D$, où $D = B^*B$ ou A^*A. Si l'on pose $\bar{z}_k = x_k$, on obtient l'équation

$$x_{k+1} = (E - \tau_{k+1}C_2)\, x_k,\ k = 0,\ 1,\ \ldots,\ x_k \in \operatorname{im} A^*, \tag{39}$$

et dans ce cas aussi $\| x_k \| = \| \bar{z}_k \|_D$, où $D = E$. Les opérateurs C_1 et C_2 sont définis dans le lemme 5.

Ainsi, dans tous les cas considérés on a obtenu l'équation de la forme

$$x_{k+1} = (E - \tau_{k+1}C)\, x_k,\ k = 0,\ 1,\ \ldots,\ x_k \in H_1 \tag{40}$$

dans le sous-espace H_1, de plus, en vertu des lemmes 4 et 5 l'opérateur C est autoadjoint dans H_1, agit dans H_1 et satisfait aux inégalités

$$\gamma_1 (x,\ x) \leqslant (Cx,\ x) \leqslant \gamma_2 (x,\ x),\ \gamma_1 > 0,\ x \in H_1, \tag{41}$$

où γ_1 et γ_2 sont empruntés des inégalités (29).

De (40) on tire

$$x_n = \prod_{j=1}^{n} (E - \tau_j C)\, x_0, \tag{42}$$

$$\| x_n \| \leqslant \| P_n (C) \| \, \| x_0 \|, \quad P_n (C) = \prod_{j=1}^{n} (E - \tau_j C).$$

Compte tenu de ce que C est autoadjoint, ainsi que des inégalités (41), il vient

$$\| P_n (C) \| \leqslant \max_{\gamma_1 \leqslant t \leqslant \gamma_2} | P_n (t) |.$$

On voit sans peine que

$$\sum_{j=1}^{n} \tau_j = - P'_n (0),$$

aussi le polynôme $P_n (t)$ est-il normé par deux conditions

$$P_n (0) = 1, \quad P'_n (0) = 0. \tag{43}$$

On aboutit donc au problème de construction d'un polynôme de degré n satisfaisant aux conditions (43) et s'écartant le moins de zéro sur le tronçon $0 < \gamma_1 \leqslant t \leqslant \gamma_2$. La construction d'un tel polynôme résout complètement le problème du choix des paramètres d'itération τ_k pour le schéma (3).

La solution exacte de ce problème nous est inconnue. On en donnera une autre solution. Comme dans la méthode des moindres résidus examinée plus haut, on maintiendra l'arbitraire dans le choix des paramètres $\tau_1,\ \tau_2,\ \ldots,\ \tau_{n-1}$, tandis que la condition $\sum_{j=1}^{n} \tau_j = 0$

sera satisfaite par le choix de τ_n suivant la formule

$$\tau_n = -\sum_{j=1}^{n-1} \tau_j.$$

De (42) on obtient l'estimation suivante:

$$\| x_n \| \leqslant \| P_{n-1}(C) \| \| E - \tau_n C \| \| x_0 \|, \quad P_{n-1}(C) = \prod_{j=1}^{n-1} (E - \tau_j C). \quad (44)$$

Choisissons maintenant les paramètres $\tau_1, \tau_2, \ldots, \tau_{n-1}$ sur la base de la condition du minimum de la norme du polynôme opératoriel $P_{n-1}(C)$. Vu qu'aucune limitation supplémentaire n'est imposée au polynôme $P_{n-1}(C)$, la solution du problème posé prend la forme (voir ch. VI, § 2):

$$\tau_k = \frac{\tau_0}{1+\rho_0\mu_k}, \quad \mu_k \in \mathfrak{M}_{n-1} = \left\{\cos\frac{(2i-1)\pi}{2(n-1)}, \ 1 \leqslant i \leqslant n-1\right\}, \quad (45)$$

$k = 1, 2, \ldots, n-1$, où sont adoptées les notations

$$\tau_0 = \frac{2}{\gamma_1+\gamma_2}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

Dans ce cas

$$P_{n-1}(t) = q_{n-1} T_{n-1}\left(\frac{1-\tau_0 t}{\rho_0}\right), \quad \| P_{n-1}(C) \| \leqslant q_{n-1}, \quad (46)$$

où $T_{n-1}(x)$ est le polynôme de Tchébychev de première espèce de degré $n-1$,

$$q_k = 2\rho_1^k/(1+\rho_1^{2k}), \quad \rho_1 = (1-\sqrt{\xi})/(1+\sqrt{\xi}).$$

Il ne reste qu'à trouver l'expression explicite de τ_n. De (46) il vient

$$\tau_n = -\sum_{j=1}^{n-1} \tau_j = P'_{n-1}(0) = -\frac{(n-1)\tau_0}{\rho_0} q_{n-1} U_{n-2}\left(\frac{1}{\rho_0}\right), \quad (47)$$

où $U_{n-2}(x)$ est le polynôme de Tchébychev de seconde espèce de degré $n-2$. On a utilisé ici la relation $T'_m(x) = mU_{m-1}(x)$. Calculons $U_{n-2}(1/\rho_0)$. Comme $\rho_0 < 1$, alors de la forme explicite de $U_{n-2}(x)$ (voir ch. I, § 4, point 2):

$$U_{n-2}(x) = \frac{(x+\sqrt{x^2-1})^{n-1} - (x+\sqrt{x^2-1})^{-(n-1)}}{2\sqrt{x^2-1}}, \quad |x| \geqslant 1,$$

on obtient finalement après des calculs simples

$$U_{n-2}\left(\frac{1}{\rho_0}\right) = \frac{1-\rho_1^{2(n-1)}}{2\rho_1^{n-1}} \frac{\rho_0}{\sqrt{1-\rho_0^2}}.$$

Portons cette expression dans (47) et il vient

$$\tau_n = -\frac{(n-1)\tau_0}{\sqrt{1-\rho_0^2}} \frac{1-\rho_1^{2(n-1)}}{1+\rho_1^{2(n-1)}}. \quad (48)$$

Compte tenu de ce que C est autoadjoint, ainsi que des inégalités (41), de la formule (48) et de l'égalité $\tau_0\gamma_2 = 1 + \rho_0$, il vient

$$\|E-\tau_n C\| \leqslant \max_{\gamma_1\leqslant t\leqslant\gamma_2} |1-\tau_n t| = 1-\tau_n\gamma_2 =$$

$$= 1+(n-1)\sqrt{\frac{1+\rho_0}{1-\rho_0}}\,\frac{1-\rho_1^{2(n-1)}}{1+\rho_1^{2(n-1)}} \leqslant 1+(n-1)\sqrt{\frac{1+\rho_0}{1-\rho_0}}. \quad (49)$$

En portant (49) et (46) dans (44), on obtient l'estimation suivante de la norme de l'erreur équivalente x_n:

$$\|x_n\| \leqslant \left(1+(n-1)\sqrt{\frac{1+\rho_0}{1-\rho_0}}\right) q_{n-1} \|x_0\|$$

à la condition que les paramètres $\tau_1, \tau_2, \ldots, \tau_n$ soient choisis suivant les formules (45) et (48).

Théorème 5. *Posons que les paramètres d'itération* τ_k, $k = 1, \ldots, n$, *pour le schéma* (3) *sont choisis suivant les formules* (45) *et* (48) *et que* $y_0 = A^*\varphi$. *On a dans ce cas pour l'erreur l'estimation*

$$\|y_n-\bar{u}\|_D \leqslant \left(1+(n-1)\sqrt{\frac{1+\rho_0}{1-\rho_0}}\right) q_{n-1} \|y_0-\bar{u}\|_D,$$

où $\bar{u}$ *est une solution normale de l'équation* (1), *tandis que* D *se définit de la façon suivante:* $D = A$, B *ou* $AB^{-1}A$, *si les conditions du lemme 4 sont remplies;* $D = B^*B$, A^*A *ou* E, *si les conditions du lemme* 5 *sont remplies. L'information à priori pour la méthode à paramètres de Tchébychev est constituée par les constantes* γ_1 *et* γ_2 *des inégalités* (29).

§ 4. Méthodes spéciales

1. Problème discret de Neumann pour l'équation de Poisson dans un rectangle. Sur l'exemple du problème mentionné montrons l'application du schéma itératif à opérateur variable B_k à la résolution de l'équation avec opérateur A dégénéré.

Supposons qu'il s'agit de trouver la solution de l'équation de Poisson dans le rectangle $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha,\ \alpha = 1, 2\}$

$$\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2} = -\varphi(x), \quad x\in G, \quad (1)$$

satisfaisant aux conditions aux limites suivantes:

$$\begin{aligned} \frac{\partial u}{\partial x_\alpha} &= -g_{-\alpha}(x_\beta), \quad x_\alpha = 0, \quad \beta = 3-\alpha, \\ -\frac{\partial u}{\partial x_\alpha} &= -g_{+\alpha}(x_\beta), \quad x_\alpha = l_\alpha, \quad \alpha = 1, 2. \end{aligned} \quad (2)$$

Sur le maillage rectangle $\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G},\ 0 \leqslant i \leqslant N_1,\ 0 \leqslant j \leqslant N_2,\ h_\alpha N_\alpha = l_\alpha,\ \alpha = 1, 2\}$ au problème (1), (2) est asso-

cié le problème de différences suivant:

$$\Lambda y = -f(x), \quad x \in \bar{\omega},$$
$$\Lambda = \Lambda_1 + \Lambda_2, \quad f(x) = \varphi(x) + \frac{2}{h_1}\varphi_1(x) + \frac{2}{h_2}\varphi_2(x), \tag{3}$$

où

$$\Lambda_\alpha y = \begin{cases} \frac{2}{h_\alpha} x_{x_\alpha}, & x_\alpha = 0, \\ y_{\bar{x}_\alpha x_\alpha}, & h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ -\frac{2}{h_\alpha} y_{\bar{x}_\alpha}, & x_\alpha = l_\alpha, \end{cases}$$
$$\varphi_\alpha(x) = \begin{cases} g_{-\alpha}(x_\beta), & x_\alpha = 0, \\ 0, & h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ g_{+\alpha}(x_\beta), & x_\alpha = l_\alpha. \end{cases} \tag{4}$$

L'espace H est composé des fonctions de mailles associées au maillage $\bar{\omega}$ avec produit scalaire $(u, v) = \sum_{x \in \bar{\omega}} u(x) v(x) \hbar_1(x_1) \hbar_2(x_2)$, où $\hbar_\alpha(x_\alpha)$ est le pas moyen,

$$\hbar(x_\alpha) = \begin{cases} h_\alpha, & h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ 0{,}5h_\alpha, & x_\alpha = 0, l_\alpha, \quad \alpha = 1, 2. \end{cases}$$

Définissons l'opérateur A comme une somme d'opérateurs A_1 et A_2, où $A_\alpha = -\Lambda_\alpha$, $\alpha = 1, 2$. Le problème (3) peut alors être écrit sous forme d'équation opératorielle

$$Au = f \tag{5}$$

avec opérateur A mentionné.

Notons les propriétés suivantes des opérateurs A_1 et A_2. Les opérateurs A_1 et A_2 sont autoadjoints dans H et permutables, autrement dit

$$A_\alpha = A_\alpha^*, \quad \alpha = 1, 2, \quad A_1 A_2 = A_2 A_1.$$

Ces propriétés permettent d'utiliser la méthode de séparation des variables et de résoudre le problème sur les valeurs propres de l'opérateur $A: Au = \lambda u$. En agissant de façon analogue que pour le problème de Dirichlet étudié en détail au point 1, § 2, ch. IV, on obtient la solution du problème sous forme de

$$\lambda_{k_1 k_2} = \lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)}, \quad \lambda_{k_\alpha}^{(\alpha)} = \frac{4}{h_\alpha^2} \sin^2 \frac{k_\alpha \pi}{2N_\alpha}, \quad k_\alpha = 0, 1, \ldots, N_\alpha,$$
$$\mu_{k_1 k_2}(i, j) = \mu_{k_1}^{(1)}(i)\, \mu_{k_2}^{(2)}(j), \quad 0 \leqslant k_\alpha \leqslant N_\alpha, \quad \alpha = 1, 2,$$
$$\mu_{k_\alpha}^{(\alpha)}(m) = \begin{cases} \sqrt{\frac{2}{l_\alpha}} \cos \frac{k_\alpha \pi m}{N_\alpha}, & 1 \leqslant k_\alpha \leqslant N_\alpha - 1, \\ \sqrt{\frac{1}{l_\alpha}} \cos \frac{k_\alpha \pi m}{N_\alpha}, & k_\alpha = 0, N_\alpha, \quad \alpha = 1, 2. \end{cases}$$

De plus, on a

$$A_\alpha \mu_{k_1k_2} = \lambda^{(\alpha)}_{k_\alpha} \mu_{k_1k_2}, \quad A\mu_{k_1k_2} = \lambda_{k_1k_2}\mu_{k_1k_2}.$$

Il en résulte que l'opérateur A possède une simple valeur propre, égale à zéro, à laquelle correspond la fonction propre $\mu_{00}(i, j) \equiv \equiv 1/\sqrt{l_1 l_2}$. Cette fonction compose la base dans le sous-espace ker A. Les fonctions $\mu_{k_1k_2}(i, j)$ pour $0 \leqslant k_\alpha \leqslant N_\alpha$ et $k_1 + k_2 \neq 0$ constituent la base du sous-espace im A.

Pour résoudre l'équation (5), considérons le schéma itératif de la méthode des directions alternées

$$\begin{gathered} B_{k+1}\frac{y_{k+1}-y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \ldots, \quad y_0 \in H, \\ B_k = (\omega_k^{(1)}E + A_1)(\omega_k^{(2)}E + A_2), \quad \tau_k = \omega_k^{(1)} + \omega_k^{(2)}. \end{gathered} \tag{6}$$

Pour ne pas imposer à $\{\tau_k\}$ et aux opérateurs $\{B_k\}$ des restrictions supplémentaires liées aux choix de la composante $\bar{y}_n \in$ im A, exigeons que le second membre f soit orthogonal à ker A. Si f ainsi fixé ne satisfait pas à cette condition, remplaçons-le dans (6) par $f_1 = = f - (f, \mu_{00})\mu_{00}$.

Notons que, pour tout k, les opérateurs B_k et A sont permutables. Aussi en vertu du corollaire du lemme 2 les conditions 1 seront-elles remplies (il faut y substituer à B l'opérateur B_k). En outre, en vertu du lemme 3, l'opérateur B_k^{-1} est une application de im A sur im A.

Profitons des faits constatés à des fins d'étude de la convergence du schéma (6). Vu que $f \in$ im A, alors, en admettant que $y_k \in$ im A, on obtient de (6) que

$$y_{k+1} = y_k - \tau_{k+1}B_{k+1}^{-1}(Ay_k - f) \in \text{im } A.$$

Aussi si l'on choisit $y_0 = 0$, $y_0 \in$ im A constituent-ils, pour tout $k \geqslant 0$, des approximations itératives de $y_k \in$ im A. Par conséquent, le schéma (6) ne peut être considéré que par rapport au sous-espace im A.

En étudiant la convergence du schéma (6) dans im A suivant la norme de l'espace H_D, on peut prendre en guise de D l'un des opérateurs E, A ou A^2. Chacun de ces opérateurs sera défini positif dans im A. Le mode d'étude de la convergence du schéma (6) est le même que celui utilisé au chapitre XI lors de la construction de la méthode des directions alternées au cas de non-dégénérescence. Aussi -imitons-nous à la formulation du problème de meilleur choix des paramètres en laissant tomber tous les calculs nécessaires.

Les meilleurs paramètres $\omega_j^{(1)}$ et $\omega_j^{(2)}$ pour le schéma (6) doivent être choisis sur la base des conditions

$$\min_{\{\omega_j\}} \max_{x,\, y\in\Omega} \left| \prod_{j=1}^{n} \frac{\omega_j^{(2)}-x}{\omega_j^{(1)}+x} \, \frac{\omega_j^{(1)}-y}{\omega_j^{(2)}+y} \right| = \rho_n,$$

où $\Omega=\Omega_1 \cup \Omega_2 \cup \Omega_3$, $\Omega_1=\{\lambda_1^{(1)} \leqslant x \leqslant \lambda_{N_1}^{(1)},\ \lambda_1^{(2)} \leqslant y \leqslant \lambda_{N_2}^{(2)}\}$,

$$\Omega_2=\{\lambda_1^{(1)} \leqslant x \leqslant \lambda_{N_1}^{(1)},\quad y=0\} \quad \text{et} \quad \Omega_3=\{x=0,\quad \lambda_1^{(2)} \leqslant y \leqslant \lambda_{N_2}^{(2)}\}.$$

Dans ce cas pour l'erreur $z_n = y_n - \bar{u}$, où $\bar{u}$ est la solution normale de l'équation (5), se justifie l'estimation

$$\| z_n \|_D \leqslant \rho_n \| z_0 \|_D.$$

Notons que pour l'exemple considéré la condition de l'orthogonalité de $\bar{u}$ à $\ker A$ s'écrit sous forme de $(u, 1) = 0$. Toute autre olution de l'équation (5) diffère de la solution normale $\bar{u}$ d'une fonction égale à une constante sur le maillage $\bar{\omega}$. Pour cette raison l'une des solutions possibles du problème (3) peut être séparée en fixant la valeur de cette solution dans l'un des nœuds du maillage $\bar{\omega}$.

Le problème des paramètres, formulé plus haut, diffère de celui étudié au § 1, ch. XI, mais peut y être réduit au moyen de quelques simplifications et au prix de la diminution de la vitesse accessible de la convergence de la méthode itérative. Notons

$$\delta=\min_{\alpha} \lambda_1^{(\alpha)}, \quad \Delta=\max_{\alpha} \lambda_{N_\alpha}^{(\alpha)}, \quad \eta=\frac{\delta}{\Delta},$$

$$\varkappa_j=\frac{1}{\Delta}\,\omega_j^{(1)}=\frac{1}{\Delta}\,\omega_j^{(2)}, \quad j=1,\ 2,\ \ldots,\ n.$$

En utilisant ces notations et la structure du domaine Ω, on est en mesure de formuler le problème du choix des paramètres ainsi: choisir $\varkappa_j$, $1 \leqslant j \leqslant n$, à partir de la condition

$$\min_{\{\varkappa_j\}} \max_{\eta \leqslant u \leqslant 1} |r_n(u, \varkappa)| = \bar{\rho}_n, \quad r_n(u, \varkappa)=\prod_{j=1}^{n} \frac{\varkappa_j-u}{\varkappa_j+u}.$$

Avec cela, apparemment, $\bar{\rho}_n > \rho_n$.

C'est justement le problème qui a été étudié au § 1 du chapitre XI. Rappelons qu'on y a obtenu les formules de $\varkappa_j$ et les nombres d'itérations $n = n_0(\varepsilon)$ qui garantissaient la réalisation de l'inégalité $\bar{\rho}_n^2 \leqslant \varepsilon$. Vu qu'il s'agit ici d'obtenir l'estimation $\bar{\rho}_n \leqslant \varepsilon$, il nous faut à cet effet substituer ε^2 à ε dans les formules de $\varkappa_j$ et $n_0(\varepsilon)$ du chapitre XI. Alors pour l'erreur de la méthode (6) se vérifiera l'estimation $\| z_n \|_D \leqslant \varepsilon \| z_0 \|_D$. Donnons l'aspect de l'estimation pour le nombre d'itérations: $n \geqslant n_0(\varepsilon)$, $n_0(\varepsilon)=\frac{1}{\pi^2} \ln \frac{4}{\eta} \ln \frac{4}{\varepsilon^2}$. A titre

d'exemple, si $l_1 = l_2 = l$ et $h_1 = h_2 = h$, on a

$$\delta = \frac{4}{h^2} \sin^2 \frac{\pi h}{2l}, \quad \Delta = \frac{4}{h^2}, \quad \eta = \sin^2 \frac{\pi h}{2}, \quad n_0(\varepsilon) = O(\ln h \ln \varepsilon).$$

Donc, pour le problème de Neumann la méthode des directions alternées, tout en ayant l'estimation du nombre d'itérations du même ordre qu'au cas du problème de Dirichlet, exige de fait deux fois plus d'itérations.

Notons que puisque les paramètres d'itération $\varkappa_j$ satisfont à l'estimation (voir § 1, ch. XI) $\eta < \varkappa_j < 1$, les paramètres $\omega_j^{(1)}$ et $\omega_j^{(2)}$ appartiennent à l'intervalle (δ, Δ). Pour cette raison, les opérateurs $\omega_k^{(\alpha)}E + A_\alpha$ sont définis positifs dans H et, pour les inverser, on peut recourir à l'algorithme ordinaire du balayage triponctuel.

2. Méthode directe pour le problème de Neumann. Voyons à présent comment s'applique à la résolution du problème de différences (3) la méthode directe, constituant une combinaison de la méthode de séparation des variables et de la méthode de réduction. Rappelons que cette méthode a été construite au point 2, § 3, ch. IV, pour le problème aux limites suivant: dans le domaine G est donnée l'équation (1), aux côtés $x_2 = 0$ et $x_2 = l_2$ sont imposées les conditions aux limites (2) et aux côtés $x_1 = 0$ et $x_1 = l_1$, au lieu des conditions de seconde espèce (2), sont imposées les conditions aux limites de troisième espèce

$$\frac{\partial u}{\partial x_1} = \varkappa_{-1} u - g_{-1}(x_1), \quad x_1 = 0,$$

$$-\frac{\partial u}{\partial x_2} = \varkappa_{+1} u - g_{+1}(x_1), \quad x_1 = l_1,$$

où $\varkappa_{-1}$ et $\varkappa_{+1}$ sont des constantes non négatives qui ne s'annulent pas en même temps. Le problème de différences correspondant ne diffère du problème (3) que par la définition de l'opérateur Λ_1. Là, on a eu affaire à l'opérateur Λ_1:

$$\Lambda_1 y = \begin{cases} \dfrac{2}{h_1}(y_{x_1} - \varkappa_{-1} y), & x_1 = 0, \\ x_{\bar{x}_1 x_1}, & h_1 \leqslant x_1 \leqslant l_1 - h_1, \\ \dfrac{2}{h_1}(-y_{\bar{x}_1} - \varkappa_{+1} y), & x_1 = l_1. \end{cases}$$

L'exigence de la non-annulation simultanée de $\varkappa_{-1}$ et $\varkappa_{+1}$ garantissait la solubilité du problème de différences et l'unicité de la solution. Dans l'algorithme de la méthode cette exigence n'était utilisée que lors de la résolution des problèmes aux limites triponctuels pour les coefficients de Fourier de la solution cherchée. C'est pourquoi, pour la résolution du problème (3), on peut formellement se servir de l'algorithme mentionné au point 2, § 3, ch. IV, en y posant $\varkappa_{-1} = \varkappa_{+1} = 0$ et de discuter séparément la question des solutions apparaissant dans les problèmes aux limites triponctuels.

Revenons au problème (3). On admettra que $f \perp \ker A$, c'est-à-dire qu'on a $(f, 1) = 0$. Le problème est alors résoluble, la solution normale $\bar{u}$ est orthogonale à $\ker A$, tandis que l'une des solutions possibles peut être séparée en fixant sa valeur dans un nœud du maillage $\bar{\omega}$. Dans l'algorithme étudié la séparation d'une des solutions possibles peut s'effectuer de façon commode, en fixant dans le nœud non pas la solution elle-même, mais un des coefficients de Fourier. Soit $y(i, j)$ la solution du problème (3). La solution normale $\bar{u}$ s'obtient alors suivant la formule

$$\bar{u} = y - (y, \mu_{00})\,\mu_{00}, \quad \mu_{00}(i, j) = 1/\sqrt{l_1 l_2}. \tag{7}$$

Fournissons à présent l'algorithme de la méthode directe de résolution du problème de Neumann (3) pour l'équation de Poisson dans un rectangle.

1) Pour $0 \leqslant i \leqslant N_1$ on calcule les valeurs de la fonction

$$\varphi(i, j) = \begin{cases} 2\,[f(i, 0) + f(i, 1)] - h_2^2\Lambda_1 f(i, 0), & j = 0, \\ f(i, 2j-1) + f(i, 2j+1) + 2f(i, 2j) - h_2^2\Lambda_1 f(i, 2j), & \\ \qquad\qquad 1 \leqslant j \leqslant M_2 - 1, & \\ 2\,[f(i, N_2) + f(i, N_2 - 1)] - h_2^2\Lambda_1 f(i, N_2), & j = M_2, \\ \varkappa_{-1} = \varkappa_{+1} = 0. & \end{cases}$$

2) Suivant l'algorithme de la transformation rapide de Fourier, on calcule les coefficients de Fourier de la fonction $\varphi(i, j)$:

$$z_{k_2}(i) = \sum_{j=0}^{M_2} \rho_j \varphi(i, j) \cos\frac{k_2\pi j}{M_2}, \quad 0 \leqslant k_2 \leqslant M_2, \quad 0 \leqslant i \leqslant N_1.$$

3) On résout les problèmes aux limites triponctuels

$$\begin{aligned} &4\sin^2\frac{k_2\pi}{2N_2}\, w_{k_2}(i) - h_2^2\Lambda_1 w_{k_2}(i) = h_2^2 z_{k_2}(i), \quad 0 \leqslant i \leqslant N_1, \\ &4\cos^2\frac{k_2\pi}{2N_2}\, y_{k_2}(i) - h_2^2\Lambda_1 y_{k_2}(i) = w_{k_2}(i), \quad 0 \leqslant i \leqslant N_1 \end{aligned} \tag{8}$$

pour $0 \leqslant k_2 \leqslant M_2$, et finalement on trouve les coefficients de Fourier $y_{k_2}(i)$ de la fonction $y(i, j)$.

4) Suivant l'algorithme de la transformation rapide de Fourier, on obtient la solution du problème sur les lignes paires du maillage $\bar{\omega}$

$$y(i, 2j) = \sum_{k_2=0}^{M_2} \rho_{k_2}\, y_{k_2}(i) \cos\frac{k_2\pi j}{M_2},$$
$$0 \leqslant j \leqslant M_2, \quad 0 \leqslant i \leqslant N_1,$$

et l'on résout les problèmes aux limites triponctuels

$$2y(i, 2j-1) - h_2^2\Lambda_1 y(i, 2j-1) =$$
$$= h_2^2 f(i, 2j-1) + u(i, 2j-2) + u(i, 2j),$$
$$0 \leqslant i \leqslant N_1, \quad 1 \leqslant j \leqslant M_2$$

pour obtenir la solution sur les lignes impaires.

On utilise ici les notations

$$M_2 = 0{,}5N_2, \quad \rho_j = \begin{cases} 1, & 1 \leqslant j \leqslant M_2 - 1, \\ 0{,}5, & j = 0,\ M_2, \end{cases}$$

l'opérateur Λ_1 est défini dans (4) et l'on admet que N_2 est la puissance de 2. Le nombre d'opérations de la méthode décrite sera égal à $O(N^2 \log_2 N)$ pour $N_1 = N_2 = N$.

La séparation d'une solution de l'ensemble des solutions du problème (3) dans l'algorithme donné se réalise de la façon suivante. De tous les problèmes aux limites triponctuels qu'il s'agit de résoudre seul le problème (8), pour $k_2 = 0$, possède une solution non unique. La séparation d'une des solutions permet de résoudre le problème posé. Le problème de différences (8) pour $k_2 = 0$ prend la forme

$$\Lambda_1 w_0(i) = -z_0(i), \quad 0 \leqslant i \leqslant N_1,$$

ou

$$\begin{aligned} (w_0)_{\bar{x}_1 x_1} &= -z_0(i), & 1 \leqslant i \leqslant N_1 - 1, \\ \frac{2}{h_1}(w_0)_{x_1} &= -z_0(0), & i = 0, \\ -\frac{2}{h_1}(w_0)_{\bar{x}_1} &= -z_0(N_1), & i = N_1. \end{aligned} \tag{9}$$

On montre sans peine, en utilisant l'orthogonalité de $f(i, j)$ à $\mu_{00}(i, j)$, que la fonction de maille $z_0(i)$ est orthogonale à la fonction $\mu_0(i) = 1/\sqrt{l_1}$ au sens de produit scalaire

$$(u, v)_1 = \sum_{x_1=0}^{l_1} u(x_1)\, v(x_1)\, \hbar_1(x_1).$$

Comme $\mu_0(i)$ est une base dans le sous-espace ker Λ_1, le problème (9) admet donc des solutions. Séparons une des solutions, en fixant la valeur $w_0(i)$ pour un certain i, $0 \leqslant i \leqslant N_1$. Posons, par exemple, $w_0(N_1) = 0$ et éliminons de (9) la condition aux limites pour $i = N_1$. Le problème de différences obtenu après une telle substitution se résout facilement par la méthode du balayage.

Après avoir trouvé l'une des solutions $y(i, j)$ du problème (3) à l'aide de l'algorithme décrit plus haut, on détermine la solution normale $\bar{u}$, si c'est nécessaire, suivant la formule (7).

Pour conclure, notons que le mécanisme analogue de séparation d'une des solutions possibles de leur ensemble peut être mis en œu-

vre également dans la méthode de réduction totale, si cette dernière est utilisée à la résolution du problème de Neumann.

3. Schémas itératifs avec opérateur B dégénéré. L'existence de méthodes directes d'inversion de l'opérateur de Laplace dans un rectangle pour le cas de conditions aux limites de Neumann autorise d'utiliser ces opérateurs en guise d'opérateur B dans des schémas itératifs implicites lors de la résolution des équations dégénérées. Comme dans ce cas l'opérateur B est dégénéré, il faut de nouveau étudier le problème de choix des paramètres d'itération.

Examinons les méthodes itératives de résolution des équations (5) avec les hypotèses suivantes: 1) l'opérateur A est autoadjoint et dégénéré; 2) le noyau de l'opérateur A est connu, c'est-à-dire que la base est donnée dans ker A ; 3) le second membre f de l'équation (5) appartient à im A, c'est-à-dire $f = \bar{f} \in$ im A. Cette condition est remplie sans peine, parce qu'on connaît la base dans ker A. La solution normale $\bar{u}$ de l'équation (5) est, dans ce cas, classique, elle satisfait à la relation

$$A\bar{u} = f. \tag{10}$$

Notons qu'en vertu du fait que l'opérateur A est autoadjoint, on a le développement orthogonal suivant de l'espace H:

$$H = \ker A \oplus \operatorname{im} A. \tag{11}$$

Pour résoudre l'équation (5), considérons le schéma implicite à deux couches

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0,\ 1,\ \ldots,\ y_0 \in H, \tag{12}$$

avec opérateur B dégénéré. Le problème consiste à trouver, à l'aide de (12), l'approximation de l'une des solutions de l'équation (5).

Formulons maintenant les hypothèses complémentaires sur les opérateurs A et B. Supposons B autoadjoint dans H et ker B = ker A. En outre, soient remplies pour tout $x \in$ im A les inégalités

$$\gamma_1 (Bx, x) \leqslant (Ax, x) \leqslant \gamma_2 (Bx, x),\ \gamma_1 > 0,\ Ax \neq 0,\ (Bx, x) > 0. \tag{13}$$

Notons qu'à partir des conditions $B = B^*$, ker B = ker A et de (11), il résulte la coïncidence de im A et im B.

Etudions le schéma (12). En accord avec (11), représentons y_k sous forme de somme

$$y_k = \bar{y}_k + \tilde{y}_k, \quad \bar{y}_k \in \operatorname{im} A, \quad \tilde{y}_k \in \ker A.$$

De (12) on obtient l'équation suivante pour y_{k+1}:

$$By_{k+1} = \varphi_k, \tag{14}$$

où $\varphi_k = By_k - \tau_{k+1} (Ay_k - f)$.

Vu que $f \in$ im A et im B = im A, on a $\varphi_k \in$ im A pour tout y_k. Par conséquent, $\varphi_k \perp$ ker B, et l'équation (14) possède un ensemble

de solutions dans le sens habituel, tandis que sa solution normale $\bar{y}_{k+1}$ satisfait à l'équation

$$B\bar{y}_{k+1} = \varphi_k. \tag{15}$$

Notons qu'en vertu de l'égalité $B\tilde{y}_k = A\tilde{y}_k = 0$ on a

$$\varphi_k = B\bar{y}_k - \tau_{k+1} (A\bar{y}_k - f). \tag{16}$$

La composante $\tilde{y}_k$ de l'approximation itérative y_k, $\tilde{y}_k \in \ker A$ n'exerce pour cette raison aucune influence sur $\bar{y}_{k+1}$. D'où il résulte qu'en résolvant l'équation (14), il suffit de trouver une solution quelconque et ce n'est qu'après l'accomplissement des itérations de calculer la projection y_n sur im A, c'est-à-dire de trouver $\bar{y}_n$.

Examinons à présent la question du choix des paramètres d'itération τ_k. En vertu de ce qui a été dit plus haut, il faut procéder au choix de manière que la suite $\bar{y}_k$ tende vers la solution normale $\bar{u}$ de l'équation (5). De (10) (15) et (16) on obtient le problème suivant pour l'erreur $\bar{z}_k = \bar{y}_k - \bar{u}$:

$$B\bar{z}_{k+1} = (B - \tau_{k+1} A)\, \bar{z}_k, \quad k = 0, 1, \ldots, \tag{17}$$

où $\bar{z}_k \in$ im A pour tout $k \geqslant 0$.

Vu que dans le sous-espace im A les opérateurs A et B, en vertu de (13), sont définis positifs, le schéma (17) peut être étudié, sous le rapport de la convergence, comme on le fait habituellement d'après la norme de l'espace énergétique H_D, où $D = A$, B ou $AB^{-1}A$. Comme dans ce cas l'opérateur $DB^{-1}A$ est autoadjoint, les paramètres τ_k peuvent être choisis suivant les formules de la méthode itérative de Tchébychev (voir § 2, ch. VI)

$$\tau_k = \frac{\tau_0}{1+\rho_0\mu_k}, \quad \mu_k \in \mathfrak{M}_n^* = \left\{\cos\frac{(2i-1)\pi}{2n}, \ 1 \leqslant i \leqslant n\right\}, \quad 1 \leqslant k \leqslant n,$$

$$\tau_0 = \frac{2}{\gamma_1+\gamma_2}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}, \tag{18}$$

$$n \geqslant n_0(\varepsilon) = \ln(0{,}5\varepsilon)/\ln \rho_1,$$

en utilisant γ_1 et γ_2 des inégalités (13). Alors pour l'erreur $\bar{z}_n$ se vérifiera, après n intérations, l'estimation

$$\|\bar{z}_n\|_D \leqslant \varepsilon \|\bar{z}_0\|_D.$$

L'idée de l'étude des méthodes itératives avec opérateur B dégénéré est la suivante. Si l'opérateur B est tel que la recherche de la solution de l'équation (14) s'avère plus simple que de l'équation de départ (5) et le rapport ξ n'est pas trop petit, ce mode de résolution approchée de l'équation (5) peut s'avérer rationnel.

Donnons l'exemple d'un problème de différences par lequel on illustrera l'application de la méthode proposée. Supposons que sur

un maillage rectangulaire

$$\bar{\omega} = \{x_{ij} = (ih_1, jh_2) \in \bar{G},\ 0 \leqslant i \leqslant N_1,\ 0 \leqslant j \leqslant N_2,\ h_\alpha N_\alpha = l_\alpha,\ \alpha = 1, 2\},$$

introduit dans un rectangle $\bar{G}$, il s'agit de rechercher la solution du problème de Neumann pour une équation elliptique à coefficients variables

$$\Lambda y = -f(x), \quad x \in \bar{\omega},$$

$$\Lambda = \Lambda_1 + \Lambda_2, \quad f(x) = \varphi(x) + \frac{2}{h_1}\varphi_1(x) + \frac{2}{h_2}\varphi_2(x), \tag{19}$$

où

$$\Lambda_\alpha y = \begin{cases} \dfrac{2}{h_\alpha} a_\alpha^{+1} y_{x_\alpha}, & x_\alpha = 0, \\ (a_\alpha y_{\bar{x}_\alpha})_{x_\alpha}, & h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ -\dfrac{2}{h_\alpha} a_\alpha y_{\bar{x}_\alpha}, & x_\alpha = l_\alpha, \end{cases}$$

$$\varphi_\alpha(x_\alpha) = \begin{cases} g_{-\alpha}(x_\beta), & x_\alpha = 0, \\ 0, & h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ g_{+\alpha}(x_\beta), & x_\alpha = l_\alpha, \end{cases}$$

$$a_1^{+1}(x) = a_1(x_1 + h_1, x_2),\ a_2^{+1}(x) = a_2(x_1, x_2 + h_2).$$

On admet que les coefficients $a_1(x)$ et $a_2(x)$ remplissent les conditions

$$0 < c_1 \leqslant a_\alpha(x) \leqslant c_2, \quad \alpha = 1, 2,\ x \in \bar{\omega}. \tag{20}$$

Le schéma (19) est l'analogue discret du problème de Neumann pour l'équation elliptique

$$\frac{\partial}{\partial x_1}\left(k_1(x)\frac{\partial u}{\partial x_1}\right) + \frac{\partial}{\partial x_2}\left(k_2(x)\frac{\partial u}{\partial x_2}\right) = -\varphi(x), \quad x \in G,$$

$$k_\alpha \frac{\partial u}{\partial x_\alpha} = -g_{-\alpha}(x_\beta), \quad x_\alpha = 0, \quad \beta = 3 - \alpha,$$

$$-k_\alpha \frac{\partial u}{\partial x_\alpha} = -g_{+\alpha}(x_\beta), \quad x_\alpha = l_\alpha, \quad \alpha = 1, 2.$$

L'espace H est défini au point 1. En introduisant l'opérateur $A = -\Lambda$, écrivons le problème de différences (19) sous forme d'équation (5). Il est aisé de vérifier que $A = A^*$, quant aux formules de différences de Green, elles donnent

$$(Ay, y) = \sum_{\alpha=1}^{2} (a_\alpha y_{\bar{x}_\alpha}^2, 1)_\alpha, \tag{21}$$

où sont utilisées les notations suivantes:

$$(u, v)_\alpha = \sum_{x_\beta=0}^{l_\beta} \sum_{x_\alpha=h_\alpha}^{l_\alpha} u(x)\, v(x)\, \hbar_\beta(x_\beta)\, h_\alpha, \quad \beta = 3-\alpha, \quad \alpha = 1, 2.$$

On montre sans peine que l'opérateur A est dégénéré et, pour tous coefficients $a_\alpha(x)$ satisfaisant à (20), le noyau de l'opérateur A est composé de fonctions de mailles constituant des constantes sur $\bar{\omega}$. Aussi, en guise de base dans ker A, peut-on choisir la fonction déjà connue $\mu_{00}(i, j) = 1/\sqrt{l_1 l_2}$.

Déterminons maintenant l'opérateur $B = -\mathring{\Lambda}$, où $\mathring{\Lambda} = \mathring{\Lambda}_1 + \mathring{\Lambda}_2$,

$$\mathring{\Lambda}_\alpha y = \begin{cases} \dfrac{2}{h_\alpha} y_{x_\alpha}, & x_\alpha = 0, \\ y_{\bar{x}_\alpha x_\alpha}, & h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ -\dfrac{2}{h_\alpha} y_{\bar{x}_\alpha}, & x_\alpha = l_\alpha, \quad \alpha = 1, 2. \end{cases}$$

L'opérateur B est autoadjoint dans H, tandis qu'au point 1 il a été noté que la base dans ker B est représentée justement par la fonction $\mu_{00}(i, j)$. Par conséquent, ker A est connu et ker A = ker B. Si, en outre, on prend la projection f sur im A et on la substitue, si nécessaire, au second membre dans le schéma (19), alors toutes les exigences imposées au schéma (12) et à l'équation (5) seront satisfaites.

Pour l'application de la méthode itérative (12), (18) il ne reste qu'à indiquer γ_1 et γ_2 dans les inégalités (13). Vu que

$$(By, y) = \sum_{\alpha=1}^{2} (y^2_{\bar{x}_\alpha}, 1)_\alpha, \tag{22}$$

tandis que les sous-espaces im A et im B coïncident, étant composés de fonctions de mailles non constantes sur le maillage $\bar{\omega}$, on obtient de (20)-(22) que $\gamma_1 = c_1$, $\gamma_2 = c_2$. L'information à priori nécessaire est ainsi trouvée.

De l'estimation (18) pour le nombre d'itérations on voit que ce dernier est indépendant du nombre d'inconnues dans le problème et ne se détermine que par le rapport c_1/c_2. Ensuite, en vertu du choix fait de l'opérateur B, l'équation (14) pour y_{k+1} se réduit au problème discret de Neumann pour l'équation de Poisson dans un rectangle. Sa solution peut être obtenue au moyen de la méthode directe exposée au point 2 en $O(N^2 \log_2 N)$ opérations arithmétiques. Le nombre total d'opérations qu'exigera la méthode proposée pour l'obtention de la solution (19) à la précision ε, sera alors égal à $Q(\varepsilon) = O(N^2 \log_2 N\, |\ln \varepsilon|)$.

CHAPITRE XIII

MÉTHODES ITÉRATIVES DE RÉSOLUTION DES ÉQUATIONS NON LINÉAIRES

On étudie dans ce chapitre les méthodes itératives de résolution des schémas aux différences non linéaires. Au § 1 on expose la théorie générale des méthodes itératives pour l'équation opératorielle abstraite non linéaire dans l'espace hilbertien avec diverses hypothèses sur l'opérateur. Le § 2 étudie l'application de la théorie générale à la résolution des analogues discrets des problèmes aux limites pour des équations quasi elliptiques de second ordre.

§ 1. Méthodes itératives. Théorie générale

1. Méthode itérative simple pour équations à opérateur monotone. On a étudié dans les chapitres précédents les méthodes itératives de résolution de l'équation opératorielle linéaire de premier ordre

$$Au = f, \tag{1}$$

donnée dans l'espace hilbertien H. La plupart des méthodes construites étaient linéaires et convergeaient à la vitesse de la progression géométrique.

Passons maintenant à l'étude des méthodes de résolution de l'équation (1) pour le cas où A est un opérateur non linéaire quelconque agissant dans H. Ce chapitre est consarcé à la construction des méthodes itératives pour la résolution des équations non linéaires (1). La construction de ces méthodes se base, en règle générale, sur l'utilisation dans les schémas itératifs implicites de l'opérateur linéaire B proche à certains égards de l'opérateur non linéaire A. Plus loin, pour diverses hypothèses sur les opérateurs A, B et D on esquissera la démonstration de théorèmes généraux sur la convergence dans H_D de la solution du schéma itératif implicite à deux couches

$$B\frac{y_{k+1}-y_k}{\tau}+Ay_k=f, \quad k=0,\ 1,\ \ldots,\ y_0\in H. \tag{2}$$

Commençons l'étude du schéma itératif (2) par le cas d'un opérateur monotone A. Rappelons que l'opérateur A donné dans l'espace hilbertien réel est appelé *monotone*, si

$$(Au-Av,\ u-v)\geqslant 0, \quad u,\ v\in H,$$

et *fortement monotone*, s'il existe un tel nombre $\delta > 0$ pour lequel avec tous $u, v \in H$

$$(Au - Av, u - v) \geqslant \delta \| u - v \|^2. \tag{3}$$

Du théorème 11, ch. V, il s'ensuit l'existence et l'unicité dans une sphère $\| u \| \leqslant \frac{1}{\delta} \| A0 - f \|$ de la solution de l'équation (1) avec opérateur fortement monotone qui, dans l'espace H, est continu et de dimension finie.

On supposera que B est un opérateur linéaire borné et défini positif dans H, tandis que D est un opérateur autoadjoint, défini positif dans H. En outre, on admet que les constantes γ_1 et γ_2 sont données dans les inégalités

$$(DB^{-1}(Au - Av), B^{-1}(Au - Av)) \leqslant \gamma_2 (DB^{-1}(Au - Av), u - v), \tag{4}$$

$$(DB^{-1}(Au - Av), u - v) \geqslant \gamma_1 (D(u - v), u - v), \tag{5}$$

avec $\gamma_1 > 0$.

L e m m e 1. *Soient remplies les conditions* (4), (5). *Alors l'équation* (1) *est résoluble de façon univalente quel que soit le second membre.*

En effet, écrivons l'équation (1) sous la forme équivalente

$$u = Su, \tag{6}$$

où l'opérateur non linéaire S est défini de la façon suivante:

$$Su = u - \tau B^{-1}Au + \tau B^{-1}f, \qquad \tau > 0.$$

Montrons que dans H_D l'opérateur S, pour $\tau < 2/\gamma_2$, est de contraction régulière, c'est-à-dire que pour tous $u, v \in H$ se justifie l'estimation

$$\| Su - Sv \|_D \leqslant \rho(\tau) \| u - v \|_D, \ \rho(\tau) < 1, \tag{7}$$

avec $\rho(\tau)$ indépendant de u et v. Dans ce cas l'assertion du lemme découlera du théorème 8, ch. V, sur les applications de contraction.

On a

$$\|Su - Sv\|^2_D = (D(Su - Sv), (Su - Sv)) = \| u - v \|^2_D - \\ - 2\tau (DB^{-1}(Au - Av), u - v) + \tau^2 (DB^{-1}(Au - Av), \\ B^{-1}(Au - Av)).$$

De (4), (5) on obtient pour $\tau < 2/\gamma_2$

$$\| Su - Sv \|^2_D \leqslant \| u - v\|^2_D - \tau(2 - \tau\gamma_2)(DB^{-1}(Au - Av), \\ u - v) \leqslant \rho^2(\tau) \| u - v \|^2_D,$$

où

$$\rho^2(\tau) = 1 - \tau(2 - \tau\gamma_2)\gamma_1. \tag{8}$$

Comme $\tau < 2/\gamma_2$, $\rho(\tau) < 1$. Le lemme est démontré.

Etudions maintenant la convergence du schéma itératif (2) dans l'hypothèse que les conditions (4), (5) sont remplies.

De (2) on tire

$$y_{k+1} = y_k - \tau B^{-1} A y_k + \tau B^{-1} f = S y_k, \qquad (9)$$

où l'opérateur non linéaire S est défini plus haut. Vu que la solution u de l'équation (1) vérifie la relation (6), on obtient de (6)-(9)

$$y_{k+1} - u = S y_k - S u, \quad k = 0, 1, \ldots,$$
$$\| y_{k+1} - u \|_D^2 = \| S y_k - S u \|_D^2 \leqslant \rho^2(\tau) \| y_k - u \|_D^2,$$

où $\rho^2(\tau)$ est défini dans (8). On voit sans peine que la meilleure estimation de la vitesse de convergence est atteinte quand $\rho(\tau)$ est minimal, c'est-à-dire pour $\tau = \tau_0 = 1/\gamma_2$. Dans ce cas $\rho_0 = \rho(\tau_0) = \sqrt{1 - \xi}$, $\xi = \gamma_1/\gamma_2$. Bref, on a démontré le théorème 1.

Théorème 1. *Supposons que les conditions* (4), (5) *sont remplies. La méthode itérative* (2) *avec* $\tau = \tau_0 = 1/\gamma_2$ *converge dans* H_D *et pour l'erreur on a l'estimation*

$$\| y_n - u \|_D \leqslant \rho_0^n \| y_0 - u \|_D, \quad \rho_0 = \sqrt{1 - \xi}, \quad \xi = \gamma_1/\gamma_2,$$

où u *est la solution de l'équation* (1). *Pour le nombre d'itérations on a l'estimation*

$$n \geqslant n_0(\varepsilon) = \ln \varepsilon / \ln \rho_0.$$

Notons que pour l'opérateur linéaire A les conditions (4), (5) peuvent être écrites de la sorte

$$(DB^{-1}Ay, B^{-1}Ay) \leqslant \gamma_2 (DB^{-1}Ay, y), \quad (DB^{-1}Ay, y) \geqslant \gamma_1 (Dy, y).$$

Par conséquent, dans ce cas elles coïncident avec les conditions imposées aux opérateurs A, B et D si l'opérateur $DB^{-1}A$ n'est pas autoadjoint dans H. La méthode construite ici passe alors à la première variante de la méthode itérative simple avec opérateur non autoadjoint (voir point 2, § 4, ch. VI).

Remarquons qu'au lieu de (4) on peut exiger que soit remplie la condition

$$\| B^{-1}(Au - Av) \|_D \leqslant \bar{\gamma}_2 \| u - v \|_D, \qquad (10)$$

qui, pour $D = B = E$, devient la condition de Lipschitz de l'opérateur A. De (10) et (15) se déduit l'inégalité (4) avec $\gamma_2 = \bar{\gamma}_2^2/\gamma_1$.

Si l'opérateur B est autoadjoint et défini positif dans H, on peut prendre en qualité d'opérateur Q l'opérateur B. Les conditions (4), (5) prennent alors la forme

$$(B^{-1}(Au - Av), Au - Av) \leqslant \gamma_2 ((Au - Av), u - v),$$
$$(Au - Av, u - v) \geqslant \gamma_1 (B(u - v), u - v), \quad \gamma_1 > 0.$$

Si B n'est ni autoadjoint ni dégénéré, alors pour $D = B^*B$ les conditions (4), (5) prennent la forme

$$(Au - Av,\ Au - Av) \leqslant \gamma_2\ (Au - Av,\ B\ (u - v)),$$

$$(Au - Av,\ B\ (u - v) \geqslant \gamma_1\ (B\ (u - v),\ B\ (u - v)),\ \gamma_1 > 0.$$

Pour $D = B = E$ la condition (5) signifie que l'opérateur A doit être fortement monotone dans H.

2. Méthodes itératives au cas d'un opérateur dérivable. On est en mesure d'améliorer l'estimation de la vitesse de convergence de la méthode itérative simple pour l'équation (1) aux dépens des limitations plus importantes imposées à l'opérateur A. A savoir, on admettra que l'opérateur A possède une dérivée Gâteau. Rappelons que l'opérateur linéaire $A'\ (u)$ est appelé *dérivée Gâteau* de l'opérateur A au point $u \in H$ si, pour tout $v \in H$, on a la relation

$$\lim_{t \to 0} \left\| \frac{A\,(u+tv) - A\,(u)}{t} - A'\,(u)\,v \right\| = 0.$$

Si l'opérateur A a une dérivée Gâteau en chaque point de l'espace H, on a l'inégalité de Lagrange

$$\|Au - Av\| \leqslant \sup_{0 \leqslant t \leqslant 1} \|A'\,(u + t\,(v - u))\|\ \|u - v\|, \quad u,\ v \in H,$$

et pour tous u, v et $w \in H$ existe un tel $t \in [0, 1]$ pour lequel

$$(Au - Av,\ w) = (A'\ (u + t\ (v - u))\ z,\ w),\ z = u - v. \quad (11)$$

Revenons à l'étude de la convergence du schéma itératif (2). On a le théorème 2.

Théorème 2. *Supposons que l'opérateur A possède dans la sphère $\Omega\ (r) = \{v: \|u - v\|_D \leqslant r\}$ une dérivée Gâteau $A'\ (v)$ qui, pour tout $v \in \Omega\ (r)$ vérifie les inégalités*

$$\begin{aligned} &(DB^{-1}A'\ (v)\ y,\ B^{-1}A'\ (v)\ y) \leqslant \gamma_2\ (DB^{-1}A'\ (v)\ y,\ y), \\ &(DB^{-1}A'(v)y,\ y) \geqslant \gamma_1\ (Dy,\ y),\ \gamma_1 > 0 \end{aligned} \quad (12)$$

pour tout $y \in H$. La méthode itérative (2) avec $\tau = 1/\gamma_2$ et $y_0 \in \Omega\ (r)$ converge dans H_D et pour l'erreur on a l'estimation

$$\|y_n - u\|_D \leqslant \rho^n\ \|y_0 - u\|_D, \quad (13)$$

où u est la solution de l'équation (1), tandis que $\rho = \sqrt{1 - \xi}$, $\xi = \gamma_1/\gamma_2$. Si l'opérateur $DB^{-1}A'\ (v)$ est autoadjoint dans H pour $v \in \Omega\ (r)$ et les inégalités

$$\gamma_1\ (Dy,\ y) \leqslant (DB^{-1}A'\ (v)\ y,\ y) \leqslant \gamma_2\ (Dy,\ y),\ \gamma_1 > 0 \quad (14$$

sont satisfaites pour tout $v \in \Omega\ (r)$ et $y \in H$, alors, avec $\tau = \tau_0 = 2/(\gamma_1 + \gamma_2)$ on a, pour le procédé itératif (2), l'estimation (13) avec $\rho = \rho_0 = (1 - \xi)/(1 + \xi)$.

En effet, de l'équation pour l'erreur

$$y_{k+1} - u = Sy_k - Su, \quad Sv = v - \tau B^{-1}\,Av + \tau B^{-1}f$$

et de l'inégalité de Lagrange, il vient

$$\| y_{k+1} - u \|_D = \| Sy_k - Su \|_D \leqslant \sup_{0 \leqslant t \leqslant 1} \| S'(v_k) \|_D \| y_k - u \|_D, \quad (15)$$

où $v_k = y_k + t(u - y_k) \in \Omega(r)$, si $y_k \in \Omega(r)$. Comme $S'(v_k) = E - \tau B^{-1}A'(v_k)$, le problème se réduit à l'estimation dans H_D de la norme de l'opérateur linéaire $E - \tau B^{-1}A'(v_k)$. De la définition de la norme de l'opérateur on tire

$$\| S'(v_k) \|_D^2 = \sup_{y \neq 0} \frac{(S'(v_k)\,y,\ S'(v_k)\,y)_D}{(y,\ y)_D} = \sup_{y \neq 0} \frac{(DS'(v_k)\,y,\ S'(v_k)\,y)}{(Dy,\ y)} =$$
$$= \sup_{z \neq 0} \frac{((E - \tau C(v_k))\,z,\ (E - \tau C(v_k))\,z)}{(z,\ z)} = \| E - \tau C(v_k) \|^2,$$

où $C(v_k) = D^{-1/2}(DB^{-1}A'(v_k))D^{-1/2}$ et l'on a fait la substitution $y = D^{-1/2} z$.

En portant la relation trouvée dans (15), on obtient

$$\| y_{k+1} - u \|_D \leqslant \sup_{0 \leqslant t \leqslant 1} \| E - \tau C(v_k) \| \, \| y_k - u \|_D.$$

De (12) on obtient que l'opérateur $C(v_k)$ vérifie pour tout $v_k \in \Omega(r)$ les inégalités

$$(C(v_k)\,y,\ C(v_k)\,y \leqslant \gamma_2 (C(v_k)\,y,\ y),$$
$$(C(v_k)\,y,\ y) \geqslant \gamma_1 (y,\ y).$$

Rappelons que l'estimation demandée pour la norme de l'opérateur linéaire $E - \tau C(v_k)$ avec les hypothèses mentionnées a été obtenue au point 2, § 4, ch. VI. A savoir, pour $\tau = 1/\gamma_2$, on a $\| E - \tau C(v_k) \| \leqslant \rho$, où $\rho = \sqrt{1 - \xi}$, $\xi = \gamma_1/\gamma_2$. La première proposition du théorème est démontrée. De façon analogue se démontre la seconde proposition. Dans ce cas l'opérateur $C(v_k)$ est autoadjoint dans H, et quant à l'estimation de la norme de l'opérateur $E - \tau C(v_k)$, elle a déjà été obtenue au point 2, § 3, ch. VI. Le théorème 2 est ainsi démontré.

Au ch. VI, outre l'estimation utilisée ici de la norme de l'opérateur $E - \tau C(v_k)$ non autoadjoint, on a obtenu une autre estimation dans l'hypothèse que sont donnés trois nombres γ_1, γ_2 et γ_3 des inégalités

$$\bar{\gamma}_1 E \leqslant C(v_k) \leqslant \bar{\gamma}_2 E, \quad \| C_1(v_k) \| \leqslant \bar{\gamma}_3, \quad \bar{\gamma}_1 > 0,$$

où $C_1 = 0{,}5\,(C - C^*)$ est la partie de symétrie gauche de l'opérateur C. Dans ce cas pour $\tau = \tau_0(1 - \varkappa\bar{\rho})$, on a l'estimation $\| E - \tau C(v_k) \| \leqslant \bar{\rho}$, où

$$\varkappa = \frac{\bar{\gamma}_3}{\sqrt{\bar{\gamma}_1\bar{\gamma}_2 + \bar{\gamma}_3^2}}, \quad \tau_0 = \frac{2}{\bar{\gamma}_1 + \bar{\gamma}_2}, \quad \bar{\rho} = \frac{1 - \bar{\xi}}{1 + \bar{\xi}}, \quad \bar{\xi} = \frac{1 - \varkappa}{1 + \varkappa} \frac{\bar{\gamma}_1}{\bar{\gamma}_2}. \quad (16)$$

Théorème 3. *Soit l'opérateur A possédant dans la sphère $\Omega(r)$ une dérivée Gâteau $A'(v)$ qui, pour tout $v \in \Omega(r)$, vérifie les inégalités*

$$\bar{\gamma}_1 (Dy, y) \leqslant (DB^{-1}A'(v)\, y, y) \leqslant \bar{\gamma}_2 (Dy, y), \quad \gamma_1 > 0, \qquad (17)$$
$$\| 0{,}5\, (DB^{-1}A'(v) - A'^*(v)\,(B^*)^{-1}D)\, y \|^2_{D^{-1}} \leqslant \bar{\gamma}_3^2 (Dy, y).$$

Alors pour $\tau = \tau_0 (1 - \varkappa\bar{\rho})$ et $y_0 \in \Omega(r)$ la méthode itérative converge dans H_D, et pour l'erreur se vérifie l'estimation (13), *où $\rho = \bar{\rho}$ est défini dans* (16).

Montrons maintenant que si l'opérateur $A'(w)$ pour $\omega \in \Omega(r)$ satisfait aux conditions (17), on a pour tous $u, v \in \Omega(r)$ les inégalités (4), (5) aux constantes $\gamma_1 = \bar{\gamma}_1$, $\gamma_2 = (\bar{\gamma}_2 + \bar{\gamma}_3)^2/\bar{\gamma}_1$. Il s'ensuit alors du lemme 1 que (1) est résoluble de façon univoque.

En vertu de (11) on a pour $u, v \in \Omega(r)$ et $t \in [0, 1]$

$$(DB^{-1}Au - DB^{-1}Av, u - v) = (Ry, y), \quad R = DB^{-1}A'(w),$$

où $y = u - v$, $w = u + t(v - u) \in \Omega(r)$. De (17) on obtient $(Ry, y) \geqslant \bar{\gamma}_1 (Dy, y)$, c'est-à-dire que l'inégalité (5) avec $\gamma_1 = \bar{\gamma}_1$ est satisfaite.

Ensuite, on a $(DB^{-1}Au - DB^{-1}Av, z) = (Ry, z)$. Représentons l'opérateur R sous forme de somme $R = R_0 + R_1$, où $R_0 = 0{,}5\,(R + R^*)$ est la partie symétrique et $R_1 = 0{,}5\,(R - R^*) = 0{,}5\,(DB^{-1}A'(w) - A'^*(w)\,(B^*)^{-1}D)$, la partie de symétrie gauche de 'opérateur R.

En vertu de l'inégalité de Cauchy-Bouniakovski et de la condition (17), on obtient

$$(R_1y, z) = (D^{-1/2}R_1y, D^{1/2}z) \leqslant (D^{-1}R_1y, R_1y)^{1/2} (Dz, z)^{1/2} =$$
$$= \| R_1y \|_{D^{-1}} (Dz, z)^{1/2} \leqslant \bar{\gamma}_3 (Dy, y)^{1/2} (Dz, z)^{1/2}.$$

De l'inégalité généralisée de Cauchy-Bouniakovski on déduit

$$(R_0y, z) \leqslant (R_0y, y)^{1/2} (R_0z, z)^{1/2} =$$
$$= (Ry, y)^{1/2} (Rz, z)^{1/2} \leqslant \bar{\gamma}_2 (Dy, y)^{1/2} (Dz, z)^{1/2}.$$

Ainsi, on a obtenu l'inégalité

$$(Ry, z) \leqslant (\bar{\gamma}_2 + \bar{\gamma}_3) (Dy, y)^{1/2} (Dz, z)^{1/2}.$$

En posant $z = B^{-1}(Au - Av)$ et en utilisant (5), il vient

$$(DB^{-1}(Au - Av), B^{-1}(Au - Av)) \leqslant \frac{(\bar{\gamma}_2 + \bar{\gamma}_3)^2}{\bar{\gamma}_1} (DB^{-1}(Au - Av), u - v).$$

La proposition est démontrée.

3. Méthode de Newton-Kantorovitch. Dans les théorèmes 2 et 3 on a admis que la dérivée Gâteau $A'(v)$ existe et satisfait aux inégalités correspondantes pour $v \in \Omega(r) = \{v : \| u - v \|_D \leqslant r\}$, où u est la solution de l'équation (1).

Il s'ensuit de la démonstration des théorèmes que pour que cela ait lieu il suffit d'exiger à chaque itération $k = 0, 1, \ldots$ la satisfaction de ces inégalités pour $v \in \Omega(r_k)$, où $r_k = \| u - y_k \|_D$.

Dans ce cas γ_1 et γ_2 (de même que $\overline{\gamma}_1$, $\overline{\gamma}_2$ et $\overline{\gamma}_3$) peuvent dépendre du numéro d'itération k. Si l'on choisit le paramètre d'itération τ suivant les formules des théorèmes 2 et 3, on obtient alors le procedé itératif non stationnaire (2) avec $\tau = \tau_{k+1}$.

De plus, on peut considérer le procédé itératif

$$B_{k+1} \frac{y_{k+1} - y_k}{\tau_{k+1}} Ay_k = f, \quad k = 0, 1, \ldots, y_0 \in H, \tag{18}$$

dont l'opérateur $B = B_{k+1}$ dépend également du numéro d'itération. Comment choisir les opérateurs B_k? Si l'opérateur A est linéaire, $A'(v) = A$ pour tout $v \in H$. Il s'ensuit alors des théorèmes 2 et 3 que pour $B = A'(v) = A$ la vitesse de convergence de la méthode itérative (2) est maximale. Notamment, pour toute approximation initiale y_0 on aura $y_1 = u$.

Choisissons à présent l'opérateur B_{k+1} au cas d'un opérateur A non linéaire de la façon suivante: $B_{k+1} = A'(y_k)$. On obtient le schéma itératif

$$A'(y_k) \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \ldots, \quad y_0 \in H. \tag{19}$$

En accord avec la terminologie adoptée, on peut dire que le procédé itératif (19) est non linéaire. Pour $\tau_k \equiv 1$, on le dénomme *méthode de Newton-Kantorovitch*. Pour apprécier la vitesse de convergence du procédé (19), on peut profiter des théorèmes 2 et 3, où à B il faut substituer $A'(y_k)$. En particulier, pour $D = E$ avec $\tau_{k+1} = 1/\gamma_2$ on a l'estimation

$$\| y_{k+1} - u \| \leqslant \rho \| y_k - u \|, \; \rho = \sqrt{1 - \gamma_1/\gamma_2} < 1, \tag{20}$$

où γ_1 et γ_2 sont empruntés aux inégalités (12) du théorème 2

$$\| (A'(y_k))^{-1} A'(v) y \|^2 \leqslant \gamma_2 ((A'(y_k))^{-1} A'(v) y, y),$$

$$((A'(y_k))^{-1} A'(v) y, y) \geqslant \gamma_1 (y, y), \; \gamma_1 > 0$$

pour $y \in H$, $v \in \Omega(r_k)$ et $r_k = \| u - y_k \|$. Il s'ensuit de (20) que $r_{k+1} = \| y_{k+1} - u \| \leqslant \rho r_k < r_k$ et, par suite $r_k \to 0$ pour $k \to \infty$. Aussi si la dérivée $A'(v)$ comme fonction de v aux valeurs comprises dans l'espace des opérateurs linéaires est continue au voisinage de la solution, alors, pour $k \to \infty$, on a $\gamma_1 \to 1$ et $\gamma_2 \to 1$. Il en résultera une accélération de la convergence de la méthode itérative (19) avec l'accroissement du numéro d'itération k.

Les raisonnements avancés montrent que les méthodes de l'aspect (19) possèdent une vitesse de convergence plus grande que celle de la progression géométrique au cas de certaines hypothèses complémentaires sur le lissage de l'opérateur $A'(v)$.

Examinons la méthode de Newton-Kantorovitch (19) avec $\tau_k \equiv 1$. Etudions la convergence de cette méthode pour les cas des hypothèses suivantes: 1) on a les inégalités

$$\| A'(v) - A'(w) \| \leqslant \alpha \| v - w \|, \quad \alpha \geqslant 0, \tag{21}$$

$$\| A'(v) y \| \geqslant \frac{1}{\beta} \| y \|, \quad y \in H, \quad \beta > 0 \tag{22}$$

vérifiées pour $v, w \in \Omega(r)$; 2) l'approximation initiale y_0 appartient à la sphère $\Omega(\bar{r})$, où $\bar{r} = \min(r, 1/(\alpha\beta))$.

Théorème 4. *Si les hypothèses* 1) *et* 2) *sont satisfaites, on a alors pour l'erreur de la méthode itérative* (19) *avec* $\tau_k \equiv 1$ *l'estimation*

$$\| y_n - u \| \leqslant \frac{1}{\alpha\beta} (\alpha\beta \| y_0 - u \|)^{2^n}. \tag{23}$$

En effet, de (19) on obtient la relation suivante:

$$A'(y_k)(y_{k+1} - u) = A'(y_k)(y_k - u) - (Ay_k - Au) = Ty_k - Tu,$$
$$Tu = A'(y_k) u - Au,$$

où u est la solution de (1). De là, en vertu de l'inégalité de Lagrange, on obtient pour l'opérateur non linéaire T

$$\| A'(y_k)(y_{k+1} - u) \| = \| Ty_k - Tu \| \leqslant \sup_{0 \leqslant t \leqslant 1} \| T'(v_k) \| \| y_k - u \|,$$

où $v_k = y_k + t(u - y_k)$. De la définition de l'opérateur T, on tire

$$T'(v_k) = A'(y_k) - A'(v_k).$$

Supposons que $y_k \in \Omega(\bar{r})$. Vu que $\bar{r} \leqslant r$, $y_k \in \Omega(r)$ et, par suite, $v_k \in \Omega(r)$. De l'inégalité (21), on tire

$$\| T'(v_k) \| = \| A'(y_k) - A'(v_k) \| \leqslant \alpha \| y_k - v_k \| = \alpha t \| y_k - u \|,$$

$$\sup_{0 \leqslant t \leqslant 1} \| T'(v_k) \| \leqslant \alpha \| y_k - u \|.$$

On obtient ainsi l'estimation

$$\| A'(y_k)(y_{k+1} - u) \| \leqslant \alpha \| y_k - u \|^2.$$

En utilisant l'inégalité (22) on en tire

$$\| y_{k+1} - u \| \leqslant \alpha\beta \| y_k - u \|^2. \tag{24}$$

Etant donné que $\| y_k - u \| \leqslant \bar{r}$ et $\alpha\beta\bar{r} \leqslant 1$, on a

$$\| y_{k+1} - u \| \leqslant \alpha\beta\bar{r} \| y_k - u \| \leqslant \| y_k - u \| \leqslant \bar{r}.$$

Par conséquent, de la condition $y_k \in \Omega(\bar{r})$ il s'ensuit que $y_{k+1} \in \Omega(\bar{r})$. Comme on a $y_0 \in \Omega(\bar{r})$, on obtient par induction que $y_k \in \Omega(\bar{r})$ pour tout $k \geqslant 0$. L'estimation (24) se justifie donc pour tout $k \geqslant 0$.

Résolvons l'inégalité (24). Multiplions-la par $\alpha\beta$ et posons $q_k = \alpha\beta \| y_k - u \|$. Pour q_k on obtient l'inégalité $q_{k+1} \leqslant q_k^2$, $k = 0, 1, \ldots$ On démontre sans peine par induction que sa solution

prend la forme $q_n \leqslant q_0^{2^n}$, $n \geqslant 0$. On a donc l'estimation

$$\alpha\beta \, \| y_n - u \| \leqslant (\alpha\beta \, \| y_0 - u \|)^{2^n}.$$

On déduit de là l'assertion du théorème.

R e m a r q u e 1. *Si l'approximation initiale y_0 est choisie de la sorte que $\bar{r} \leqslant \rho/(\alpha\beta)$, $\rho < 1$, il s'ensuit de* (23) *l'estimation*

$$\| y_n - u \| \leqslant \rho^{2^n - 1} \, \| y_0 - u \|$$

et l'estimation

$$n \geqslant n_0 (\varepsilon) = \log_2 (\ln \varepsilon / \ln \rho + 1)$$

pour le nombre d'itérations.

R e m a r q u e 2. *Si au lieu de la condition* (21) *est satisfaite l'inégalité*

$$\| A' (v) - A' (w) \| \leqslant \alpha \, \| v - w \|^p, \quad p \in (0, 1],$$

on a alors pour l'erreur l'estimation

$$\| y_n - u \| \leqslant \frac{1}{\sqrt[p]{\alpha\beta}} \left(\sqrt[p]{\alpha\beta} \, \| y_0 - u \| \right)^{(p+1)^n},$$

$$\left(\bar{r} = \min \left(r, \frac{1}{\sqrt[p]{\alpha\beta}} \right) \right).$$

Avec la démonstration du théorème 4 on a obtenu l'estimation pour l'erreur (23). Cette estimation, du point de vue de son application pratique, ne présente pas d'intérêt, mais elle acquiert de l'importance pour la théorie de la méthode, puisqu'elle montre comment se réalise la convergence près de la solution u.

Le théorème 4 permet de distinguer les domaines où la solution n'existe pas. De fait, le théorème postule que y_k converge vers u si $\| y_0 - u \| \leqslant \bar{r}$. Donc, si les itérations ne convergent pas, il n'y aura pas de solutions de l'équation (1) pour la sphère $\| y_0 - v \| \leqslant \bar{r}$ de centre au point y_0.

Notons que si l'opérateur A possède pour la sphère $\Omega (r)$ une seconde dérivée Gâteau, on aura dans l'inégalité (21)

$$\alpha = \sup_{0 \leqslant t \leqslant 1} \| A'' (v + t (w - v)) \|.$$

Avec la mise en œuvre du schéma itératif (19) il faut, pour chaque k, résoudre l'équation opératorielle linéaire

$$A' (y_k) \, v = F (y_k), \tag{25}$$

où

$$F (y_k) = A' (y_k) \, y_k - \tau_{k+1} (A y_k - f). \tag{26}$$

Si v est une solution précise de l'équation (25), alors dans (19) $y_{k+1} = v$.

L'opérateur $A'(y_k)$ doit être calculé à chaque itération, et cela peut exiger de laborieuses opérations. Voyons un exemple. Soit A l'opérateur correspondant à un système d'équations non linéaires

$$\varphi_i(u) = 0, \quad i = 1, 2, \ldots, m, \quad u = (u_1, u_2, \ldots, u_m).$$

La dérivée Gâteau $A'(y)$ au point $y = (y_1, y_2, \ldots, y_m)$ est une matrice carrée à éléments $a_{ij}(y)$, où

$$a_{ij}(y) = \left.\frac{\partial \varphi_i(u)}{\partial u_j}\right|_{u=y}, \quad i, j = 1, 2, \ldots, m.$$

Par conséquent, à chaque itération il faut calculer m^2 éléments de la matrice $A'(y)$, tandis que le nombre d'inconnues dans le problème vaut m.

Pour éviter le calcul de la dérivée $A'(y_k)$ à chaque itération, on utilise le schéma (19) modifié suivant:

$$A'(y_{km}) \frac{y_{km+i+1} - y_{km+i}}{\tau_{km+i+1}} + Ay_{km+i} = f,$$

$$i = 0, 1, \ldots, m-1, \quad k = 0, 1, \ldots$$

La dérivée A' est ici calculée après toutes les m itérations et est utilisée pour la recherche des approximations intermédiaires y_{km+1}, $y_{km+2}, \ldots, y_{(k+1)m}$. Pour $m = 1$, on obtient le schéma itératif (19).

4. Méthodes itératives à deux étapes. Le schéma itératif (19) peut être utilisé de façon rationnelle au cas où l'opérateur $A'(y_k)$ est facilement inversible. Dans ce cas la solution précise v de l'équation (25) est prise pour une nouvelle approximation itérative y_{k+1} qui satisfait au schéma (19). On obtient ainsi le schéma itératif dont l'opérateur B_{k+1} est donné sous forme explicite: $B_{k+1} = A'(y_k)$

Si l'équation (25) est résolue de façon approchée, au moyen, par exemple, d'une méthode itérative auxiliaire (interne) et en guise de y_{k+1} on prend la m-ième approximation itérative v_m, alors y_{k+1} satisfait au schéma général (18) avec un certain $B_{k+1} \neq A'(y_k)$. Dans ce cas l'aspect explicite de l'opérateur B_{k+1} n'est pas utilisé et la connaissance de sa structure n'est nécessaire que pour l'étude de la convergence du schéma itératif (18). Les méthodes itératives construites de cette façon sont quelquefois appelées à deux étapes, en entendant par cette expression l'existence d'un algorithme spécial d'inversion de l'opérateur B_{k+1}.

Décrivons d'une manière détaillée le schéma général de construction des méthodes à deux étapes. Supposons que pour la résolution de l'équation linéaire (25) on utilise une certaine méthode itérative à deux couches

$$\bar{B}_{n+1} \frac{v_{n+1} - v_n}{\omega_{n+1}} + A'(y_k) v_n = F(y_k), \quad n = 0, 1, \ldots, m-1, \quad (27)$$

où $F(y_k)$ est défini dans (26), $\{\omega_n\}$ étant le jeu de paramètres d'itération, $\bar{B}_{n+1}$, des opérateurs dans H qui peuvent dépendre de y_k, tandis que $v_0 = y_k$.

Exprimons v_m au moyen de y_k. Cherchons d'abord l'équation pour l'erreur $z_n = v_n - v$, où v est la solution de l'équation (25). De (25) et (27) on tire

$$z_{n+1} = S_{n+1}z_n, \quad n = 0, 1, \ldots, \quad S_n = E - \omega_n \bar{B}_n^{-1} A'(y_k)$$

et, par suite,

$$z_m = v_m - v = T_m z_0 = T_m (v_0 - v), \quad T_m = S_m S_{m-1} \ldots S_1, \qquad v_m = (E - T_m) v + T_m y_k. \tag{28}$$

A partir de (25), (26), on obtient

$$v = [A'(y_k)]^{-1} F(y_k) = y_k - \tau_{k+1} [A'(y_k)]^{-1} (Ay_k - f).$$

En portant v trouvé dans (28), il vient

$$y_{k+1} = v_m = y_k - \tau_{k+1} (E - T_m) [A'(y_k)]^{-1} (Ay_k - f).$$

Il en suit que y_{k+1} satisfait au schéma itératif (18) si l'on pose

$$B_{k+1} = A'(y_k) (E - T_m)^{-1}. \tag{29}$$

De cette façon, la mise en œuvre de un pas de la méthode à deux étapes consiste dans le calcul de $F(y_k)$ suivant la formule (26) et l'exécution de m itérations suivant le schéma (27) avec l'approximation initiale $v_0 = y_k$. L'approximation obtenue v_m est prise en guise de y_{k+1}.

Examinons le schéma itératif (18), (29). Pour apprécier la vitesse de convergence, on peut utiliser les théorèmes 2 et 3 dans lesquels B est remplacé par B_{k+1} et τ par τ_{k+1}. L'incovénient de ce choix du paramètre τ est la nécessité d'apprécier γ_1, γ_2 et γ_3 de façon suffisamment précise.

Notons que pour la construction de la méthode à deux étapes on aurait pu se référer non pas à l'équation (25), mais à l'équation qui en est « proche »

$$Rv = F(y_k),$$

où l'opérateur linéaire R est en quelque sorte équivalent à l'opérateur $A'(y_k)$. On a dans ce cas dans le schéma itératif (18)

$$B_{k+1} \equiv B = R (E - T_m)^{-1}.$$

Etudions ce procédé en détail. Soient remplies les conditions

$$R = R^* > 0, \quad T_m R = R T_m, \tag{30}$$

$$\| T_m \|_R \leqslant q < 1. \tag{31}$$

L e m m e 2. *Admettons que les conditions* (30), (31) *sont remplies. L'opérateur* $B = R(E - T_m)^{-1}$ *est alors autoadjoint et défini*

positif dans H et on a les inégalités

$$(1 - q) B \leqslant R \leqslant (1 + q) B. \quad (32)$$

Examinons l'opérateur $B^{-1} = (E - T_m) R^{-1}$. A partir de (30), cherchons $(E - T_m^*) R = R(E - T_m)$ ou $R^{-1} (E - T_m^*) = (E - T_m) R^{-1}$. L'opérateur B^{-1} est donc autoadjoint dans H.

Puisqu'en vertu de (30) l'opérateur T_m est autoadjoint dans H_R, on a

$$\| T_m \|_R = \sup_{x \neq 0} \frac{|(T_m x, x)_R|}{(x, x)_R} = \sup_{x \neq 0} \frac{|(RT_m x, x)|}{(Rx, x)} \leqslant q < 1.$$

Par conséquent, pour tout $x \in H$, on a l'inégalité

$$|(RT_m x, x)| \leqslant q (Rx, x).$$

En posant ici $x = R^{-1} y$, il vient

$$|(T_m R^{-1} y, y)| \leqslant q (R^{-1} y, y),$$

aussi pour $y \in H$ obtient-on

$$(1 - q) (R^{-1} y, y) \leqslant ((E - T_m) R^{-1} y, y) \leqslant (1 + q) (R^{-1} y, y).$$

Ainsi, on a obtenu l'estimation

$$(1 - q) R^{-1} \leqslant B^{-1} \leqslant (1 + q) R^{-1}. \quad (33)$$

Vu que R^{-1} et B^{-1} sont des opérateurs autoadjoints dans H et $q < 1$, il s'ensuit alors du lemme 9, § 1, ch. V, que les inégalités (33) et (32) sont équivalentes. Le lemme est démontré.

Lemme 3. *Supposons que l'opérateur A possède dans la sphère* $\Omega (r)$ *une dérivée Gâteau* $A' (v)$ *qui, pour tout* $v \in \Omega (r)$, *satisfait aux inégalités*

$$c_1 (Ry, y) \leqslant (A' (v) y, y) \leqslant c_2 (Ry, y), \quad c_1 > 0, \quad (34)$$

$$\| 0{,}5 [A' (v) - (A' (v))^*] y \|^2_{R^{-1}} \leqslant c_3^2 (Ry, y), \quad c_3 \geqslant 0, \quad (35)$$

et que soient remplies les conditions (30), (31). *Alors se vérifient les inégalités* (17) *du théorème* 3, *où*

$$\bar{\gamma}_1 = c_1 (1 - q), \quad \bar{\gamma}_2 = c_2 (1 + q), \quad \bar{\gamma}_3 = c_3 (1 + q)^2,$$
$$D = B = R (E - T_m).$$

En effet, en vertu du lemme 2 l'opérateur D est autoadjoint et défini positif dans H. D'autre part les inégalités (17), pour $D = B$, prennent la forme

$$\bar{\gamma}_1 (By, y) \leqslant (A' (v) y, y) \leqslant \bar{\gamma}_2 (By, y), \quad (36)$$

$$\| 0{,}5 [A' (v) - (A' (v))^*] y \|^2_{B^{-1}} \leqslant \bar{\gamma}_3^2 (By, y). \quad (37)$$

Les inégalités (36), avec $\bar{\gamma}_1$ et $\bar{\gamma}_2$ mentionnés dans le lemme 3, s'ensuivent de (32) et (34), tandis que (37) se déduit de (32), (33)

et (35), car

$$\| z \|^2_{B^{-1}} = (B^{-1}z, z) \leqslant (1 + q)(R^{-1}z, z) = (1 + q) \| z \|^2_{R^{-1}},$$
$$(Rz, z) \leqslant (1 + q)(Bz, z).$$

En utilisant le lemme 3, on peut démonter l'analogue du théorème 3 pour la méthode à deux étapes.

Théorème 5. *Admettons que les conditions du lemme 3 sont remplies, et la méthode à deux étapes est construite sur la base de l'équation $Rv = F(y_k)$ avec utilisation de l'opérateur résolvant T_m. Si dans le schéma itératif* (18) *avec $B_{k+1} \equiv B = R(E - T_m)^{-1}$, décrivant cette méthode à deux étapes, on choisit $\tau_k \equiv \tau_0(1 - \varkappa\bar{\rho})$ et $y_0 \in \Omega(r)$, alors pour l'erreur se vérifie l'estimation*

$$\| y_n - u \|_B \leqslant \bar{\rho}^n \| y_0 - u \|_B,$$

où u est la solution de l'équation (1); $\bar{\rho}$, $\varkappa$ *et* τ_0 *sont définis dans* (16) *avec* $\bar{\gamma}_1$, $\bar{\gamma}_2$ *et* $\bar{\gamma}_3$ *donnés dans le lemme* 3.

5. Autres méthodes itératives. Dans ce point on va donner une description sommaire de quelques méthodes itératives qu'on utilise également pour la résolution de l'équation (1) possédant un opérateur A non linéaire.

Soit $\Phi(u)$ la fonctionnelle dans H dérivable suivant Gâteau. L'opérateur A agissant dans H est dit *potentiel* s'il existe une fonctionnelle dérivable $\Phi(u)$ telle que $Au = \operatorname{grad} \Phi(u)$, quel que soit u. Le gradient de la fonctionnelle $\Phi(u)$ se définit ici par l'égalité $\frac{d}{dt}\Phi(u + tv)|_{t=0} = (\operatorname{grad} \Phi(u), v)$. A titre d'exemple d'opérateur potentiel on peut indiquer l'opérateur A borné, linéaire et autoadjoint agissant dans l'espace hilbertien H. Il est engendré par la fonctionnelle $\Phi(u) = 0{,}5\,(Au, u)$.

Supposons que l'opérateur A est continûment dérivable dans H. L'opérateur A est potentiel seulement et rien que seulement quand la dérivée Gâteau $A'(v)$ est un opérateur autoadjoint dans H.

Si l'opérateur A est potentiel, la formule

$$\Phi(u) = \int_0^1 (A(u_0 + t(u - u_0)), u - u_0)\, dt,$$

où u_0 est un élément quelconque mais fixé de H, fournit le procédé de construction de la fonctionnelle $\Phi(u)$ suivant l'opérateur A.

Si l'opérateur A est engendré par le gradient d'une fonctionnelle strictement convexe, la dérivée $A'(v)$ est un opérateur défini positif dans H pour tout $v \in H$. Dans ce cas, pour obtenir la solution approchée de l'équation (25), on peut recourir aux méthodes itératives du type variationnel, c'est ainsi, par exemple, que dans (27) les paramètres d'itération ω_{n+1} doivent être choisis suivant les formules des méthodes de la plus grande pente, des moindres résidus, etc.

Voyons, en guise d'exemple, la méthode à deux étapes (18), (29) pour laquelle $\tau_{k+1} \equiv 1$, tandis que dans le schéma (27) $m = 1$ et $\overline{B}_1 = E$. Alors $B_{k+1} = E/\omega_1$. Si pour le procédé itératif auxiliaire (27) le paramètre ω_1 est choisi suivant les formules de la méthode des moindres résidus (ou des moindres corrections), on obtient alors (voir points 2, 3, § 2, ch. VIII)

$$\omega_1 = \frac{(A'(y_k)\, r_k,\ r_k)}{\| A'(y_k)\, r_k \|^2}, \quad r_k = Ay_k - f. \tag{38}$$

Dans ce cas la méthode itérative à deux étapes se décrit par la formule

$$\frac{y_{k+1} - y_k}{\omega_1} + Ay_k = f, \quad k = 0,\ 1,\ \ldots, \tag{39}$$

où ω_1 est défini dans (38).

Dans la situation où l'opérateur A n'est pas potentiel, le paramètre ω_1 peut être choisi suivant les formules de la méthode des moindres erreurs, en posant dans (27) $\overline{B}_1 = [(A'(y_k))^*]^{-1}$ et

$$\omega_1 = \frac{(r_k,\ r_k)}{\| (A'(y_k))^*\, r_k \|^2}, \quad r_k = Ay_k - f. \tag{40}$$

Dans ce cas la méthode à deux étapes prend la forme

$$\frac{y_{k+1} - y_k}{\omega_1} + (A'(y_k))^*\, Ay_k = (A'(y_k))^* f, \quad k = 0,\ 1,\ \ldots, \tag{41}$$

où ω_1 est défini dans (40).

On voit sans peine que dans la méthode (38), (39) le paramètre ω_1 est choisi sur la base de la condition du minimum de $\| A'(y_k)(y_{k+1} - y_k) + Ay_k - f \|$, tandis que dans la méthode (40), (41) il est choisi sur la base de la condition du minimum de la norme $\| y_{k+1} - y_k + [A'(y_k)]^{-1} (Ay_k - f) \|$.

Le problème de la résolution de l'équation $Au = f$ au cas d'un opérateur potentiel peut parfois être remplacé par le problème de minimisation de la fonctionnelle engendrant cet opérateur. Notons qu'il existe toujours un procédé simple de transformation du problème de résolution de l'équation (1) en un problème de minimisation, même si l'opérateur A n'est pas potentiel.

En effet soit $\Phi(u)$ la fonctionnelle donnée dans H et présentant un point minimum unique $u = 0$. En guise d'exemple d'une telle fonctionnelle on peut fournir $\Phi(u) = (Du, u)$, où D est un opérateur autoadjoint défini positif dans H. Ensuite, pour l'équation (1) considérée étudions la fonctionnelle

$$F(u) = \Phi(Au - f), \quad u \in H.$$

Si l'équation (1) a une solution u, elle fournit apparemment un minimum à la fonctionnelle $F(u)$.

Décrivons la méthode de minimisation de la fonctionnelle (méthode de la descente). Supposons que l'équation (1) est engendrée par le gradient de la fonctionnelle strictement convexe $\Phi(u)$. La suite minimisante est posée construite suivant le schéma itératif (19), autrement dit suivant la formule

$$y_{k+1} = y_k - \tau_{k+1} [A'(y_k)]^{-1} (Ay_k - f), \quad k = 0, 1, \ldots \tag{42}$$

Posons

$$w_k = [A'(y_k)]^{-1} \operatorname{grad} \Phi(y_k), \tag{43}$$

où, en vertu des hypothèses faites, $\operatorname{grad} \Phi(y_k) = Ay_k - f$. Ecrivons (42) sous la forme

$$y_{k+1} = y_k - \tau_{k+1} w_k.$$

Notons que l'opérateur $A'(y_k)$ est défini positif et autoadjoint dans H. Ensuite, à partir de la définition de la dérivée Gâteau de la fonctionnelle on a

$$\lim_{\tau_{k+1} \to 0} \left[\frac{\Phi(y_k - \tau_{k+1} w_k) - \Phi(y_k)}{\tau_{k+1}} \right] + (\operatorname{grad} \Phi(y_k), \ w_k) = 0.$$

Vu que $A'(y_k) w_k = \operatorname{grad} \Phi(y_k)$, on a

$$(\operatorname{grad} \Phi(y_k), \ w_k) = (A'(y_k) w_k, \ w_k) > 0.$$

Par conséquent, il existe un tel $\tau_{k+1} > 0$ pour lequel $\Phi(y_{k+1})$ sera strictement inférieur à $\Phi(y_k)$.

Si la suite minimisante $\{y_k\}$ est construite suivant le schéma explicite (18) ($B_k \equiv E$), c'est-à-dire suivant les formules

$$y_{k+1} = y_k - \tau_{k+1} (Ay_k - f),$$

le passage de y_k à y_{k+1} s'effectue suivant la direction du gradient de la fonctionnelle $\Phi(u)$ au point y_k. Ces méthodes sont généralement appelées *méthodes de descente par gradient*. Il existe des algorithmes de choix des paramètres d'itération τ_k, toutefois on ne s'arrêtera pas sur ces questions ici.

Fournissons, en conclusion, la généralisation de la méthode explicite des gradients conjugués, qui est utilisée pour la minimisation de la fonctionnelle avec les hypothèses posées plus haut. Les formules de l'algorithme de Flesher-Rievs ont la forme

$$\begin{aligned} y_{k+1} &= y_k - a_{k+1} w_k, & k &= 0, 1, \ldots, \\ w_k &= \operatorname{grad} \Phi(y_k) + b_k w_{k-1}, & k &= 1, 2, \ldots, \\ w_0 &= \operatorname{grad} \Phi(y_0), \end{aligned}$$

où

$$b_k = \frac{\| \operatorname{grad} \Phi(y_k) \|^2}{\| \operatorname{grad} \Phi(y_{k-1}) \|^2}, \quad k = 1, 2, \ldots,$$

quant au paramètre a_{k+1}, il est choisi sur la base de la condition du minimum de $\Phi(y_k - a_{k+1}w_k)$. Ce problème de recherche du minimum de la fonction à une variable se résout par l'une des méthodes de l'analyse numérique.

§ 2. Méthodes de résolution des schémas aux différences non linéaires

1. Schéma aux différences pour une équation quasi linéaire elliptique unidimensionnelle. La théorie générale des méthodes itératives exposée au § 1 sera appliquée pour la recherche de la solution approchée des schémas aux différences elliptiques non linéaires. Commençons par des exemples les plus simples.

Examinons le troisième problème aux limites pour une équation quasi linéaire unidimensionnelle sous la forme divergente

$$Lu = \frac{d}{dx} k_1\left(x, u, \frac{du}{dx}\right) - k_0\left(x, u, \frac{du}{dx}\right) = -\varphi(x), \quad 0 \leqslant x \leqslant l,$$

$$k_1\left(x, u, \frac{du}{dx}\right) = \varkappa_0(u) - \mu_0, \quad x = 0, \tag{1}$$

$$-k_1\left(x, u, \frac{du}{dx}\right) = \varkappa_1(u) - \mu_1, \quad x = l.$$

On supposera que les fonctions $k_1(x, p_0, p_1)$, $k_0(x, p_0, p_1)$, $\varkappa_0(p_0)$ et $\varkappa_1(p_0)$ sont continues en p_0 et p_1 et que les conditions de l'ellipticité sont satisfaites

$$\sum_{\alpha=0}^{1} [k_\alpha(x, p_0, p_1) - k_\alpha(x, q_0, q_1)](p_\alpha - q_\alpha) \geqslant c_1 \sum_{\alpha=0}^{1} (p_\alpha - q_\alpha)^2, \tag{2}$$

$$[\varkappa_\alpha(p_0) - \varkappa_\alpha(q_0)](p_0 - q_0) \geqslant 0, \quad \alpha = 0, 1, \tag{3}$$

où $c_1 > 0$ est une constante positive, $0 \leqslant x \leqslant l$, $|p_0|$, $|q_0|$, $|p_1|$, $|q_1| < \infty$.

Sur un maillage régulier $\bar\omega = \{x_i = ih, i = 0, 1, \ldots, N, hN = l\}$ mettons en accord avec le problème (1) le schéma aux différences

$$\Lambda y_i = -f_i, \quad 0 \leqslant i \leqslant N, \tag{4}$$

où

$$f_i = \begin{cases} \varphi(0) + \frac{2}{h}\mu_0, & i = 0, \\ \varphi(x_i), & 1 \leqslant i \leqslant N-1, \\ \varphi(l) + \frac{2}{h}\mu_1, & i = N. \end{cases}$$

L'opérateur de différences Λ se détermine à l'aide des formules:

$$\Lambda y_i = \frac{1}{2}\{[k_1(x, y, y_x)]_{\bar{x}} + [k_1(x, y, y_{\bar{x}})]_x - k_0(x, y, y_x) - \\ - k_0(x, y, y_{\bar{x}})\}_i, \quad 1 \leqslant i \leqslant N-1,$$

$$\Lambda y_0 = \frac{1}{h}[k_1(0, y_0, y_{x,0}) + k_1(h, y_1, y_{\bar{x},1})] - \\ - k_0(0, y_0, y_{x,0}) - \frac{2}{h}\varkappa_0(y_0), \quad i=0,$$

$$\Lambda y_N = -\frac{1}{h}[k_1(l-h, y_{N-1}, y_{x,N-1}) + k_1(l, y_N, y_{\bar{x},N})] - \\ - k_0(l, y_N, y_{\bar{x},N}) - \frac{2}{h}\varkappa_1(y_N), \quad i=N.$$

Si dans l'espace $H = H(\bar{\omega})$ on définit l'opérateur non linéaire A par la relation $A = -\Lambda$, le schéma aux différences (4) s'écrira alors sous la forme d'une équation opératorielle $Au = f$.

Etudions les propriétés de l'opérateur non linéaire A agissant de H dans H. Rappelons que le produit scalaire dans $H(\bar{\omega})$ se définit par la formule

$$(u, v) = \sum_{i=1}^{N-1} u_i v_i h + 0{,}5h(u_0 v_0 + u_N v_N),$$

tandis qu'au moyen de $(u, v)_{\omega^+}$ et $(u, v)_{\omega^-}$ se notent les sommes

$$(u, v)_{\omega^+} = \sum_{i=1}^{N} u_i v_i h, \quad (u, v)_{\omega^-} = \sum_{i=0}^{N-1} u_i v_i h,$$

de sorte que

$$(u, v) = \frac{1}{2}[(u, v)_{\omega^+} + (u, v)_{\omega^-}].$$

Montrons qu'avec la satisfaction des conditions (2), (3) l'opérateur A est fortement monotone dans $H(\bar{\omega})$, c'est-à-dire qu'est vérifiée l'inégalité

$$(Au - Av, u - v) \geqslant c_1 \| u - v \|^2, \quad c_1 > 0, \tag{5}$$

où c_1 est défini dans (2).

Posons $p_0 = \bar{p}_0 = u_i$, $q_0 = \bar{q}_0 = v_i$, $p_1 = u_{x,i}$, $\bar{p}_1 = u_{\bar{x},i}$, $q_1 = v_{x,i}$, $\bar{q}_1 = v_{\bar{x},i}$. En utilisant la définition de l'opérateur A, les formules de sommation par parties (voir (7), (9), § 2, ch. V) et les condi-

tions (2), (3), on obtient

$$(Au-Av,\ u-v)=(\Lambda v-\Lambda u,\ u-v)=$$

$$=\frac{1}{2}\sum_{i=0}^{N}h\left\{\sum_{\alpha=0}^{1}[k_\alpha(x,\bar p_0,\bar p_1)-k_\alpha(x,\bar q_0,\bar q_1)](\bar p_\alpha-\bar q_\alpha)\right\}_i+$$

$$+\frac{1}{2}\sum_{i=0}^{N-1}h\left\{\sum_{\alpha=0}^{1}[k_\alpha(x,p_0,p_1)-k_\alpha(x,q_0,q_1)](p_\alpha-q_\alpha)\right\}_i+$$

$$+[\varkappa_1(\bar p_0)-\varkappa_1(\bar q_0)](\bar p_0-\bar q_0)\,|_{i=N}+[\varkappa_0(p_0)-\varkappa_0(q)](p_0-q_0)\,|_{i=0}\geqslant$$

$$\geqslant\frac{c_1}{2}\sum_{i=1}^{N}h\sum_{\alpha=0}^{1}(\bar p_\alpha-\bar q_\alpha)_i^2+\frac{c_1}{2}\sum_{i=0}^{N-1}h\sum_{\alpha=0}^{1}(p_\alpha-q_\alpha)_i^2.$$

Compte tenu de l'égalité $u_{x,\,i}=u_{\bar x,\,i+1}$, écrivons l'estimation obtenue sous la forme

$$(Au-Av,\ u-v)\geqslant\frac{c_1}{2}\Big[(u-v,\ u-v)_{\omega^+}+(u-v,\ u-v)_{\omega^-}+$$

$$+\sum_{i=1}^{N}h\,(u-v)^2_{\bar x,\,i}+\sum_{i=0}^{N-1}h\,(u-v)^2_{x,\,i}\Big]=$$

$$=c_1\,[\,\|u-v\|^2+((u-v)^2_{\bar x},\ 1)_{\omega^+}]\geqslant c_1\,\|u-v\|^2.$$

De la remarque 2 au lemme 12, ch. V, il s'ensuit que cette estimation ne peut être améliorée.

Ainsi, on a établi une forte monotonie de l'opérateur A. En vertu de la continuité des fonctions $k_\alpha(x,\ p_0,\ p_1)$ et $\varkappa_\alpha(p_0)$, l'opérateur A est continu dans H. Il s'ensuit donc du théorème 11, ch. V que la solution de l'équation $Au=f$, et, partant, du problème de différences (4), dans la sphère $\|u\|\leqslant\frac{1}{c_1}\|A0-f\|$ existe et est unique.

Si $k_\alpha(x,\ p_0,\ p_1)$ et $\varkappa_\alpha(p_0)$, $\alpha=0,\ 1$ sont des fonctions constamment dérivables de leurs arguments, on peut au lieu de (2), (3) utiliser d'autres conditions suffisantes garantissant une forte monotonie de l'opérateur A.

Supposons que soient satisfaites les conditions

$$c_1\sum_{\alpha=0}^{1}\xi_\alpha^2\leqslant\sum_{\alpha,\,\beta=0}^{1}a_{\alpha\beta}(x,\ p_0,\ p_1)\,\xi_\alpha\xi_\beta\leqslant c_2\sum_{\alpha=0}^{1}\xi_\alpha^2,\qquad c_1>0,\tag{6}$$

$$0\leqslant\sigma_\alpha(p_0)\leqslant c_3,\quad \alpha=1,\ 2,\tag{7}$$

où $\xi=(\xi_0,\ \xi_1)$ est un vecteur quelconque et

$$a_{\alpha\beta}(x,\ p_0,\ p_1)=\frac{\partial k_\alpha(x,\ p_0,\ p_1)}{\partial p_\beta},\quad \sigma_\alpha(p_0)=\frac{\partial\varkappa_\alpha(p_0)}{\partial p_0},\quad \alpha,\ \beta=0,\ 1.$$

Montrons que des conditions (6), (7) se déduisent (2), (3). En effet, on a les égalités

$$k_\alpha(x, p_0, p_1) - k_\alpha(x, q_0, q_1) =$$

$$= \int_0^1 \frac{d}{dt} k_\alpha(x, tp_0 + (1-t) q_0, tp_1 + (1-t) q_1)\, dt =$$

$$= (p_0 - q_0) \int_0^1 \frac{\partial k_\alpha(x, s_0, s_1)}{\partial s_0}\, dt + (p_1 - q_1) \int_0^1 \frac{\partial k_\alpha(x, s_0, s_1)}{\partial s_1}\, dt =$$

$$= \sum_{\beta=0}^{1} (p_\beta - q_\beta) \int_0^1 a_{\alpha\beta}(x, s_0, s_1)\, dt, \quad \alpha = 0, 1,$$

où $s_0 = tp_0 + (1-t) q_0$, $s_1 = tp_1 + (1-t) q_1$. En multipliant cette égalité par $p_\alpha - q_\alpha$ et en la sommant en α de 0 à 1, on obtient, compte tenu de (6),

$$\sum_{\alpha=0}^{1} [k_\alpha(x, p_0, p_1) - k_\alpha(x, q_0, q_1)] (p_\alpha - q_\alpha) =$$

$$= \int_0^1 \sum_{\alpha, \beta=0}^{1} a_{\alpha\beta}(x, s_0, s_1)(p_\alpha - q_\alpha)(p_\beta - q_\beta)\, dt \geqslant$$

$$\geqslant c_1 \int_0^1 \sum_{\alpha=0}^{1} (p_\alpha - q_\alpha)^2\, dt = c_1 \sum_{\alpha=0}^{1} (p_\alpha - q_\alpha)^2.$$

On a ainsi obtenu l'inégalité (2). De façon analogue, de (7) on déduit l'inégalité (3)

$$[\varkappa_\alpha(p_0) - \varkappa_\alpha(q_0)] (p_0 - q_0) = \int_0^1 \frac{\partial \varkappa_\alpha(s_0)}{\partial s_0}\, dt\, (p_0 - q_0)^2 \geqslant 0.$$

Les conditions (6), (7) garantissent donc l'existence et l'unicité de la solution du problème de différences (4).

Cherchons maintenant la dérivée Gâteau de l'opérateur A en supposant que les fonctions $k_\alpha(x, p_0, p_1)$ et $\varkappa_\alpha(p_0)$, $\alpha = 0, 1$ possèdent des dérivées bornées en p_0 et p_1 d'ordre exigé.

A partir de la définition de la dérivée Gâteau d'un opérateur non linéaire il vient

$$A'(u)\,y_i = -\frac{1}{2}\{[a_{11}(x, u, u_x)\,y_x]_{\bar{x},i} + [a_{11}(x, u, u_{\bar{x}})\,y_{\bar{x}}]_{x,i} + $$
$$+ [a_{10}(x, u, u_x)\,y]_{\bar{x},i} + [a_{10}(x, u, u_{\bar{x}})\,y]_{x,i}\} +$$
$$+ \frac{1}{2}\{a_{01}(x, u, u_x)\,y_{x,i} + a_{01}(x, u, u_{\bar{x}})\,y_{\bar{x},i} +$$
$$+ [a_{00}(x, u, u_x) + a_{00}(x, u, u_{\bar{x}})]\,y_i\}, \quad 1 \leqslant i \leqslant N-1.$$

Pour $i=0$, on obtient

$$A'(u)\,y_0 = -\frac{1}{h}[a_{11}(0, u_0, u_{x,0}) + a_{11}(h, u_1, u_{\bar{x},1}) -$$
$$- ha_{01}(0, u_0, u_{x,0}) + ha_{10}(h, u_1, u_{\bar{x},1})]\,y_{x,0} +$$
$$+ \frac{2}{h}\Big[\sigma_0(u_0) - \frac{1}{2}a_{10}(0, u_0, u_{x,0}) - \frac{1}{2}a_{10}(h, u_1, u_{\bar{x},1}) +$$
$$+ \frac{h}{2}a_{00}(0, u_0, u_{x,0})\Big]\,y_0,$$

tandis que pour $i=N$, on aura

$$A'(u)\,y_N = \frac{1}{h}[a_{11}(l-h, u_{N-1}, u_{x,N-1}) + a_{11}(l, u_N, u_{\bar{x},N}) +$$
$$+ ha_{01}(l, u_N, u_{\bar{x},N}) - ha_{10}(l-h, u_{N-1}, u_{x,N-1})]\,y_{\bar{x},N} +$$
$$+ \frac{2}{h}\Big[\sigma_1(u_N) + \frac{1}{2}a_{10}(l, u_N, u_{\bar{x},N}) + \frac{1}{2}a_{10}(l-h, u_{N-1}, u_{x,N-1}) +$$
$$+ \frac{h}{2}a_{00}(l, u_N, u_{\bar{x},N})\Big]\,y_N.$$

Notons qu'avec le calcul de $A'(u)\,y_0$ et de $A'(u)\,y_N$ on a utilisé les relations

$$y_1 = y_0 + hy_{x,0}, \quad y_{N-1} = y_N - hy_{\bar{x},N}. \tag{8}$$

Etudions les propriétés de la dérivée Gâteau $A'(u)$ de l'opérateur A.

L e m m e 4. *Si sont remplies les conditions*

$$\frac{\partial k_1(x, p_0, p_1)}{\partial p_0} = \frac{\partial k_0(x, p_0, p_1)}{\partial p_1}, \tag{9}$$

$A'(u)$ est alors un opérateur autoadjoint dans H. Avec la satisfaction des conditions (6), (7) *il devient défini positif dans H.*

En effet, en utilisant les formules de sommation par parties ainsi que les relations (8), on obtient

$$(A'(u)\,y,\ z)=\frac{1}{2}\sum_{i=0}^{N-1} h\,[a_{11}(x,\ u,\ u_x)\,y_x z_x + a_{10}(x,\ u,\ u_x)\times$$

$$\times\, y z_x + a_{01}(x,\ u,\ u_x)\,y_x z + a_{00}(x,\ u,\ u_x)\,yz]_i +$$

$$+\frac{1}{2}\sum_{i=1}^{N} h\,[a_{11}(x,\ u,\ u_{\bar{x}})\,y_{\bar{x}} z_{\bar{x}} + a_{10}(x,\ u,\ u_{\bar{x}})\times$$

$$\times\, y z_{\bar{x}} + a_{01}(x,\ u,\ u_{\bar{x}})\,y_{\bar{x}} z + a_{00}(x,\ u,\ u_x)\,yz]_i +$$

$$+\sigma_0(u_0)\,y_0 z_0 + \sigma_1(u_N)\,y_N z_N. \qquad (10)$$

En comparant cette expression à l'expression de $(y,\ A'(u)\,z)$, on obtient que si la condition $a_{10}(x,\ p_0,\ p_1) = a_{01}(x,\ p_0,\ p_1)$, qui est une autre forme d'écriture de (9), est satisfaite, l'opérateur $A'(u)$ est autoadjoint dans H pour tout $u \in H$.

Supposons à présent que ce sont les conditions (6), (7) qui sont remplies. En posant dans (10) $z_i \equiv y_i$, il vient

$$(A'(u)\,y,\ y)\geqslant\frac{c_1}{2}\Big[\sum_{i=0}^{N-1} h\,(y_i^2+y_{x,\,i}^2)+\sum_{i=1}^{N} h\,(y_i^2+y_{\bar{x},\,i}^2)\Big]=$$

$$=c_1\,[(y,\ y)+(y_{\bar{x}}^2,\ 1)_{\omega^+}]\geqslant c_1\,(y,\ y), \qquad (11)$$

c'est-à-dire que l'opérateur $A'(u)$ est défini positif dans H. Le lemme est démontré.

Notons qu'en vertu du théorème 2, ch. V, il s'ensuit du fait que la dérivée Gâteau de l'opérateur continu A est définie positive, que ce dernier est fortement monotone. Donc, une fois les conditions (6), (7) remplies, l'opérateur A est fortement monotone.

En posant dans (10) $z_i \equiv y_i$, on obtient en vertu des conditions (6), (7) l'estimation supérieure

$$(A'(u)\,y,\ y)\leqslant\frac{c_2}{2}\Big[\sum_{i=0}^{N-1} h\,(y_i^2+y_{x,\,i}^2)+\sum_{i=1}^{N} h\,(y_i^2+y_{\bar{x},\,i}^2)\Big]+$$

$$+\,c_3\,(y_0^2+y_N^2)=c_2\,[(y,\ y)+(y_{\bar{x}}^2,\ 1)_{\omega^+}]+c_3\,(y_0^2+y_N^2).$$

A partir de l'inégalité (36) du lemme 15, ch. V, pour $\varepsilon = 1$, on obtient que

$$y_0^2+y_N^2\leqslant c_4\,[(y,\ y)+(y_{\bar{x}}^2,\ 1)_{\omega^+}],\qquad c_4=\frac{8+l^2}{l\sqrt{16+l^2}}. \qquad (12)$$

On a donc

$$(A'(u)\,y,\ y)\leqslant \gamma_2\,[(y,\ y)+(y_{\bar{x}}^2,\ 1)_{\omega^+}],\quad \gamma_2=c_2+c_3c_4. \qquad (13)$$

Définissons dans l'espace $H = H(\bar{\omega})$ l'opérateur linéaire R, application de H sur H, suivant les formules

$$Ry_i = \begin{cases} -\frac{2}{h} y_{x,0} + y_0, & i = 0, \\ -y_{\bar{x}x,i} + y_i, & 1 \leqslant i \leqslant N-1, \\ \frac{2}{h} y_{\bar{x},N} + y_N, & i = N. \end{cases}$$

De la première formule de différences de Green, il vient

$$(Ry, y) = (y, y) + (y_{\bar{x}}^2, 1)_{\omega^+}. \tag{14}$$

Alors de (11), (13), (14) il s'ensuit facilement qu'avec la satisfaction des conditions (6), (7) on a pour la dérivée Gâteau $A'(u)$ de l'opérateur A les inégalités

$$\gamma_1 (Ry, y) \leqslant (A'y, y) \leqslant \gamma_2 (Ry, y), \tag{15}$$

où $\gamma_1 = c_1 > 0$, $\gamma_2 = c_2 + c_3 c_4$, c'est-à-dire que les opérateurs R et A' sont énergétiquement équivalents aux constantes qui ne dépendent pas du pas h du maillage.

Rappelons qu'on a obtenu plus haut l'inégalité

$$(Au - Av, u - v) \geqslant c_1 [\| u - v \|^2 + ((u - v)_{\bar{x}}^2, 1)_{\omega^+}]$$

au cas où les conditions (2), (3) sont remplies. De là et de (14) il résulte que si les conditions (2), (3) sont satisfaites, on obtient l'estimation

$$(Au - Av, u - v) \geqslant \gamma_1 (R(u - v), u - v), \quad \gamma_1 = c_1 > 0. \tag{16}$$

Montrons maintenant que pour tous $u, v \in H$, on a l'estimation

$$(R^{-1}(Au - Av), Au - Av) \leqslant \gamma_2 (Au - Av, u - v), \tag{17}$$

où $\gamma_2 = c_2 (1 + c_4)$ si sont satisfaites les conditions

$$\sum_{\alpha=0}^{1} [k_\alpha (x, p_0, p_1) - k_\alpha (x, q_0, q_1)]^2 \leqslant$$

$$\leqslant c_2 \sum_{\alpha=0}^{1} [k_\alpha (x, p_0, p_1) - k_\alpha (x, q_0, q_1)] (p_\alpha - q_\alpha), \tag{18}$$

$$[\varkappa_\alpha (p_0) - \varkappa_\alpha (q_0)]^2 \leqslant c_2 [\varkappa_\alpha (p_0) - \varkappa_\alpha (q_0)] (p_0 - q_0).$$

En effet, pour démontrer (17), il suffit d'obtenir pour tous $u, v, z \in H$ l'estimation

$$(Au - Av, z)^2 \leqslant \gamma_2 (Au - Av, u - v) (Rz, z). \tag{19}$$

Dans ce cas, en posant ici $z = R^{-1}(Au - Av)$, on aboutit à (17).

Posons:

$$p_0=\bar p_0=u_i,\quad q_0=\bar q_0=v_i,\quad s_0=\bar s_0=z_i,$$
$$p_1=u_{x,i},\quad \bar p_1=u_{\bar x,i},\quad q_1=v_{x,i},\quad \bar q_1=v_{\bar x,i},$$
$$s_1=z_{x,i},\quad \bar s_1=z_{\bar x,i}.$$

En profitant de la définition de l'opérateur A et, compte tenu des formules de sommation par parties, il vient

$$(Au-Av,\ z)^2=(\Lambda v-\Lambda u,\ z)^2=$$
$$=\Big\{\frac{1}{2}\sum_{\alpha=0}^{1}([k_\alpha(x,\ p_0,\ p_1)-k_\alpha(x,\ q_0,\ q_1)],\ s_\alpha)_{\omega^-}+$$
$$+\frac{1}{2}\sum_{\alpha=0}^{1}([k_\alpha(x,\ \bar p_0,\ \bar p_1)-k_\alpha(x,\ \bar q_0,\ \bar q_1)],\ \bar s_\alpha)_{\omega^+}+$$
$$+[\varkappa_1(\bar p_0)-\varkappa_1(\bar q_0)]\,\bar s_0\,|_{i=N}+[\varkappa_0(p_0)-\varkappa_0(q_0)]\,s_0\,|_{i=0}\Big\}^2.$$

En utilisant l'inégalité de Cauchy-Bouniakovski, on obtient successivement

$$(Au-Av,\ z)^2\leqslant\Big\{\frac{1}{2}\sum_{\alpha=0}^{1}([k_\alpha(x,\ p_0,\ p_1)-$$
$$-k_\alpha(x,\ q_0,\ q_1)]^2,\ 1)_{\omega^-}^{1/2}(s_\alpha^2,\ 1)_{\omega^-}^{1/2}+$$
$$+\frac{1}{2}\sum_{\alpha=0}^{1}([k_\alpha(x,\ \bar p_0,\ \bar p_1)-k_\alpha(x,\ \bar q_0,\ \bar q_1)]^2,\ 1)_{\omega^+}^{1/2}\times$$
$$\times(\bar s_\alpha^2,\ 1)_{\omega^+}^{1/2}+[\varkappa_1(\bar p_0)-\varkappa_1(\bar q_0)]\,\bar s_0\,|_{i=N}+$$
$$+[\varkappa_0(p_0)-\varkappa_0(q_0)]\,s_0\,|_{i=0}\Big\}^2\leqslant$$
$$\leqslant\Big\{\frac{1}{2}\sum_{\alpha=0}^{1}([k_\alpha(x,\ p_0,\ p_1)-k_\alpha(x,\ q_0,\ q_1)]^2,\ 1)_{\omega^-}+$$
$$+\frac{1}{2}\sum_{\alpha=0}^{1}([k_\alpha(x,\ \bar p_0,\ \bar p_1)-k_\alpha(x,\ \bar q_0,\ \bar q_1)]^2,\ 1)_{\omega^+}+$$
$$+[\varkappa_1(\bar p_0)-\varkappa_1(\bar q_0)]_{i=N}^2+[\varkappa_0(p_0)-\varkappa_0(q_0)]_{i=0}^2\Big\}\times$$
$$\times\Big\{\frac{1}{2}\sum_{\alpha=0}^{1}[(s_\alpha^2,\ 1)_{\omega^-}+(\bar s_\alpha^2,\ 1)_{\omega^+}]+\bar s_0^2\,|_{i=N}+s_0^2\,|_{i=0}\Big\}.$$

Etant donné que l'égalité

$$(Au-Av,\ u-v)=$$

$$=\frac{1}{2}\sum_{\alpha=0}^{1}([k_\alpha(x,\ p_0,\ p_1)-k_\alpha(x,\ q_0,\ q_1)](p_\alpha-q_\alpha),\ 1)_{\omega^-}+$$

$$+\frac{1}{2}\sum_{\alpha=0}^{1}([k_\alpha(x,\ \bar p_0,\ \bar p_1)-k_\alpha(x,\ \bar q_0,\ \bar q_1)](\bar p_\alpha-\bar q_\alpha),\ 1)_{\omega^+}+$$

$$+[\varkappa_1(\bar p_0)-\varkappa_1(\bar q_0)](\bar p_0-\bar q_0)\,|_{i=N}+[\varkappa_0(p_0)-\varkappa_0(q_0)](p_0-q_0)\,|_{i=0}$$

est vraie, et qu'en vertu de (12), (14) et des notations introduites

$$\frac{1}{2}\sum_{\alpha=0}^{1}[(s_\alpha^2,\ 1)_{\omega^-}+(\bar s_\alpha^2,\ 1)_{\omega^+}]+\bar s_0^2\,|_{i=N}+s_0^2\,|_{i=0}=$$

$$=\frac{1}{2}[(z^2,\ 1)_{\omega^+}+(z^2,\ 1)_{\omega^-}+(z_{\bar x}^2,\ 1)_{\omega^+}+(z_x^2,\ 1)_{\omega^-}]+z_N^2+z_0^2=$$

$$=(z^2,\ 1)+(z_{\bar x}^2,\ 1)_{\omega^+}+z_N^2+z_0^2\leqslant(1+c_4)(Rz,\ z),$$

on obtient l'estimation (19), si les conditions (18) sont remplies. La proposition est démontrée.

2. Méthode itérative simple. Examinons maintenant les méthodes itératives de résolution du schéma aux différences non linéaire (4) qu'on a construit. Supposons, au préalable, que les conditions (2), (3) et (18) sont satisfaites.

Pour résoudre l'équation (4), recourrons à la méthode itérative simple du type implicite

$$B\frac{y_{k+1}-y_k}{\tau}+Ay_k=f,\quad k=0,\ 1,\ \ldots,\quad y_0\in H,\tag{20}$$

où $A=-\Lambda$, $B=R$ (l'opérateur R étant défini plus haut). Il s'ensuit de (20) que pour trouver y_{k+1}, y_k étant donné, il faut résoudre l'équation linéaire

$$By_{k+1}=\varphi,\quad \varphi=By_k-\tau(Ay_k-f)$$

ou sous forme développée

$$-y_{k+1}(i-1)+cy_{k+1}(i)-y_{k+1}(i+1)=h^2\varphi(i),\quad 1\leqslant i\leqslant N-1,$$

$$cy_{k+1}(0)-2y_{k+1}(1)=h^2\varphi(0),\qquad i=0,$$

$$-2y_{k+1}(N-1)+cy_{k+1}(N)=h^2\varphi(N),\quad i=N,$$

où $c=2+h^2$. Vu que $c>2$, le problème discret aux limites peut être résolu par la méthode du balayage monotone en $O(N)$ opérations arithmétiques.

Il reste à indiquer le rôle du paramètre d'itération τ et de fournir l'estimation du nombre d'itérations exigées. Les conditions (2), (3)

et (18) étant remplies, on a les estimations (16) et (17) qui peuvent être écrites sous la forme

$$(Au - Av,\ u - v) \geqslant \gamma_1 (B(u - v),\ u - v), \quad \gamma_1 = c_1 > 0.$$
$$(B^{-1}(Au - Av),\ Au - Av) \leqslant \gamma_2 (Au - Av,\ u - v),$$
$$\gamma_2 = c_2 (1 + c_4), \quad (21)$$

où c_1 est donné dans (2), c_2 dans (18) et c_4 dans (12).

Comme l'opérateur B est autoadjoint et défini positif, la convergence de la méthode (20) sera étudiée dans l'espace énergétique H_D, où $D = B$. Avec le choix considéré de l'opérateur D les inégalités (21) coïncident avec les inégalités (4), (5). Aussi peut-on profiter pour le choix du paramètre d'itération τ du théorème 1. On obtient que pour $\tau = 1/\gamma_2 = 1/(c_2 (1 + c_4))$ la méthode itérative (20) converge dans H_D, et pour l'erreur on a l'estimation $\| y_n - u \|_B \leqslant \leqslant \rho^n \| y_0 - u \|_B$, $\rho = \sqrt{1 - \xi}$, $\xi = \gamma_1/\gamma_2$ pour toute approximation initiale y_0.

Donc, si les conditions (2), (3), (18) sont satisfaites, la méthode itérative simple (20) avec la valeur du paramètre τ mentionnée permet d'obtenir la solution du schéma aux différences non linéaire (4) avec la précision ε en $n \geqslant n_0(\varepsilon)$ itérations, où

$$n_0(\varepsilon) = \frac{\ln \varepsilon}{\ln \rho} = \frac{2 \ln \varepsilon}{\ln \left(1 - \frac{c_1}{c_2 (1 + c_4)}\right)}.$$

Vu que les constantes c_1, c_2 et c_4 ne dépendent pas du pas h du maillage, le nombre d'itérations $n_0(\varepsilon)$ n'est fonction que de ε et ne varie pas avec la dégénérescence du maillage.

Examinons maintenant la méthode itérative (20) dans l'hypothèse de la satisfaction de (6), (7) pour les dérivées $a_{\alpha\beta} = \partial k_\alpha / \partial p_\beta$ et $\sigma_\alpha = \partial \varkappa_\alpha / \partial p_0$, ainsi que de la condition de symétrie (9). Alors pour la dérivée Gâteau de l'opérateur A se vérifieront les inégalités (15) qui, en vertu du choix de $B = R$, prennent la forme

$$\gamma_1 (By,\ y) \leqslant (A'(v)\, y,\ y) \leqslant \gamma_2 (By,\ y), \quad v,\ y \in H, \qquad (22)$$

où $\gamma_1 = c_1$, $\gamma_2 = c_2 + c_3 c_4$, c_1, c_2 et c_3 étant définis dans (6), (7), tandis que c_4 l'est dans (12).

Soit $D = B$. L'opérateur $DB^{-1}A'(v)$, égal à $A'(v)$, sera, en vertu du lemme 4, autoadjoint dans H et, par conséquent, les conditions du théorème 2 sont satisfaites, tandis que les inégalités (22) se ramènent aux inégalités (14). Le paramètre τ dans le schéma (20) doit donc être pris égal à $\tau = \tau_0 = 2/(\gamma_1 + \gamma_2)$. En outre, pour l'erreur $y_n - u$ et pour le nombre d'itérations se vérifieront les estimations

$$\| y_n - u \|_B \leqslant \rho_0^n \| y_0 - u \|, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2} = \frac{c_1}{c_2 + c_3 c_4},$$
$$n \geqslant n_0(\varepsilon) = \ln \varepsilon / \ln \rho_0.$$

Ici, comme pour la méthode précédente, le nombre d'itérations est indépendant du pas h du maillage. Pour l'opérateur B, choisi en vertu de la première formule de différences de Green, on aura la représentation suivante pour la norme $\| z \|_B$:

$$\| z \|_B^2 = (z, z) + (z_{\bar{x}}^2, 1)_{\omega^+}.$$

On a examiné les méthodes de résolution du schéma aux différences non linéaire approximant l'équation unidimensionnelle quasi linéaire sur un maillage régulier. Il est aisé d'étendre ces études au cas de maillage irrégulier quelconque ainsi qu'à des schémas aux différences approximant les principaux problèmes aux limites pour l'équation elliptique quasi linéaire de second ordre dans un rectangle.

3. Méthodes itératives pour équations aux différences elliptiques quasi linéaires dans un rectangle. Dans le rectangle $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha, \alpha = 1, 2\}$ à frontière Γ il s'agit de trouver la solution de l'équation

$$\sum_{\alpha=1}^{2} \frac{\partial}{\partial x_\alpha} k_\alpha \left(x, u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}\right) - k_0 \left(x, u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}\right) = -\varphi(x), \quad x \in G, \tag{23}$$

qui satisfait aux conditions aux limites de troisième espèce

$$\begin{gathered} k_\alpha \left(x, u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}\right) = \varkappa_{-\alpha}(x, u) - g_{-\alpha}(x), \ x_\alpha = 0, \\ -k_\alpha \left(x, u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}\right) = \varkappa_{+\alpha}(x, u) - g_{+\alpha}(x), \ x_\alpha = l_\alpha, \ \alpha = 1, 2. \end{gathered} \tag{24}$$

Supposons, comme dans le cas unidimensionnel, que les conditions suivantes sont satisfaites. Les fonctions $k_\alpha(x, p)$ et $\varkappa_{\pm\alpha}(p_0)$ sont continues en $p = (p_0, p_1, p_2)$ et p_0, et, en outre,

$$\sum_{\alpha=0}^{2} [k_\alpha(x, p) - k_\alpha(x, q)](p_\alpha - q_\alpha) \geqslant c_1 \sum_{\alpha=0}^{2} (p_\alpha - q_\alpha)^2, \quad c_1 > 0,$$

$$\sum_{\alpha=0}^{2} [k_\alpha(x, p) - k_\alpha(x, q)]^2 \leqslant c_2 \sum_{\alpha=0}^{2} [k_\alpha(x, p) - k_\alpha(x, q)](p_\alpha - q_\alpha),$$

$$[\varkappa_{\pm\alpha}(p_0) - \varkappa_{\pm\alpha}(q_0)]^2 \leqslant c_2 [\varkappa_{\pm\alpha}(p_0) - \varkappa_{\pm\alpha}(q_0)](p_0 - q_0), \quad \alpha = 1, 2,$$

où $c_1 > 0$ et $c_2 > 0$, $x \in \bar{G}$ et $|p|, |q| < \infty$.

Introduisons dans les domaines $\bar{G}$ un maillage régulier rectangulaire

$$\bar{\omega} = \{x_{ij} = (ih_1, jh_2), \quad 0 \leqslant i \leqslant N_1, \quad 0 \leqslant j \leqslant N_2, \quad h_\alpha N_\alpha = l_\alpha, \quad \alpha = 1, 2\}.$$

Le schéma aux différences du type le plus simple, correspondant au problème (23), (24), prend la forme

$$\Lambda y = -f, \qquad x \in \bar{\omega}, \tag{25}$$
$$\Lambda = \Lambda_1 + \Lambda_2, \quad f = \varphi + 2\varphi_1/h_1 + 2\varphi_2/h_2,$$

où

$$\varphi_\alpha(x) = \begin{cases} g_{-\alpha}(x), & x_\alpha = 0, \\ 0, & h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ g_{+\alpha}(x), & x_\alpha = l_\alpha,\ 0 \leqslant x_{3-\alpha} \leqslant l_{3-\alpha}, \end{cases}$$

tandis que les opérateurs Λ_α, $\alpha = 1, 2$, sont définis par les formules :

1) pour $h_\beta \leqslant x_\beta \leqslant l_\beta - h_\beta$, on a

$$\Lambda_\alpha y = \frac{1}{2}\{[k_\alpha(x, y, y_{\bar{x}_1}, y_{\bar{x}_2})]_{x_\alpha} + [k_\alpha(x, y, y_{x_1}, y_{x_2})]_{\bar{x}_\alpha}\} -$$
$$- \frac{1}{4}[k_0(x, y, y_{\bar{x}_1}, y_{\bar{x}_2}) + k_0(x, y, y_{x_1}, y_{x_2})], \quad h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha;$$

$$\Lambda_\alpha y = \frac{1}{h_\alpha}[k_\alpha^{+1_\alpha}(x, y, y_{\bar{x}_1}, y_{\bar{x}_2}) + k_\alpha(x, y, y_{x_1}, y_{x_2})] -$$
$$- \frac{1}{2}k_0(x, y, y_{x_1}, y_{x_2}) - \frac{2}{h_\alpha}\varkappa_{-\alpha}(x, y), \quad x_\alpha = 0;$$

$$\Lambda_\alpha y = -\frac{1}{h_\alpha}[k_\alpha(x, y, y_{\bar{x}_1}, y_{\bar{x}_2}) + k_\alpha^{-1_\alpha}(x, y, y_{x_1}, y_{x_2})] -$$
$$- \frac{1}{2}k_0(x, y, y_{\bar{x}_1}, y_{\bar{x}_2}) - \frac{2}{h_\alpha}\varkappa_{+\alpha}(x, y), \quad x_\alpha = l_\alpha;$$

2) pour $x_\beta = 0$, on a

$$\Lambda_\alpha y = [k_\alpha(x, y, y_{x_1}, y_{x_2})]_{\bar{x}_\alpha} - \frac{1}{2}k_0(x, y, y_{x_1}, y_{x_2})$$

avec $h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha$;

$$\Lambda_\alpha y = \frac{2}{h_\alpha}k_\alpha(x, y, y_{x_1}, y_{x_2}) - k_0(x, y, y_{x_1}, y_{x_2}) - \frac{2}{h_\alpha}\varkappa_{-\alpha}(x, y)$$

avec $x_\alpha = 0$;

$$\Lambda_\alpha y = -\frac{2}{h_\alpha}k_\alpha^{-1_\alpha}(x, y, y_{x_1}, y_{x_2}) - \frac{2}{h_\alpha}\varkappa_{+\alpha}(x, y), \ x_\alpha = l_\alpha;$$

3) pour $x_\beta = l_\beta$ on a

$$\Lambda_\alpha y = [k_\alpha(x, y, y_{\bar{x}_1}, y_{\bar{x}_2})]_{x_\alpha} - \frac{1}{2}k_0(x, y, y_{\bar{x}_1}, y_{\bar{x}_2})$$

avec $h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha$;

$$\Lambda_\alpha y = \frac{2}{h_\alpha}k_\alpha^{+1_\alpha}(x, y, y_{\bar{x}_1}, y_{\bar{x}_2}) - \frac{2}{h_\alpha}\varkappa_{-\alpha}(x, y), \ x_\alpha = 0;$$

$$\Lambda_\alpha y = -\frac{2}{h_\alpha}k_\alpha(x, y, y_{\bar{x}_1}, y_{\bar{x}_2}) - k_0(x, y, y_{\bar{x}_1}, y_{\bar{x}_2}) - \frac{2}{h_\alpha}\varkappa_{+\alpha}(x, y)$$

avec $x_\alpha = l_\alpha$.

Dans le cas concerné $\beta = 3-\alpha$, $\alpha = 1, 2$, et on utilise les notations

$$k_1^{+1_1}(x, y, y_{\bar{x}_1}, y_{\bar{x}_2})|_{x_{ij}} = \\ = k_1(x_{i+1,j}, y(i+1, j), y_{\bar{x}_1}(i+1, j), y_{\bar{x}_2}(i+1, j)),$$

et des notations analogues pour $k_1^{-1_1}$ et $k_2^{\pm 1_2}$.

Définissons dans l'espace H des fonctions de mailles associées à $\bar{\omega}$ le produit scalaire

$$(u, v) = \sum_{i=0}^{N_1} \sum_{j=0}^{N_2} \hbar_1(i) \hbar_2(j) u(i, j) v(i, j),$$

$$\hbar_\alpha(k) = \begin{cases} h_\alpha, & 1 \leqslant k \leqslant N_\alpha - 1, \\ 0{,}5h_\alpha, & k = 0, N_\alpha \end{cases}$$

et les opérateurs $A_\alpha = -\Lambda_\alpha$, $\alpha = 1, 2$, $A = A_1 + A_2$, $R = R_1 + R_2$, où

$$R_\alpha y = \begin{cases} -\frac{2}{h_\alpha} y_{x_\alpha} + \frac{1}{2} y, & x_\alpha = 0, \\ -y_{\bar{x}_\alpha x_\alpha} + \frac{1}{2} y, & h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha, \\ \frac{2}{h_\alpha} y_{\bar{x}_\alpha} + \frac{1}{2} y, & x_\alpha = l_\alpha, \ \alpha = 1, 2, \end{cases}$$

et $0 \leqslant x_\beta \leqslant l_\beta$. Dans ce cas le schéma aux différences (25) s'écrira sous la forme d'une équation opératorielle

$$Au = f \tag{26}$$

avec opérateur non linéaire A.

En utilisant les hypothèses formulées plus haut sur les coefficients $k_\alpha(x, p)$ et $\varkappa_{\pm\alpha}(p_0)$, on obtient, comme dans le cas unidimensionnel, les inégalités (16) et (17), où c_4 est une constante de l'inégalité

$$\sum_{j=0}^{N_2} \hbar_2(j)[y^2(0, j) + y^2(N_1, j)] + \sum_{i=0}^{N_1} \hbar_1(i)[y^2(i, 0) + y^2(i, N_2)] \leqslant \\ \leqslant c_4 \Big[\sum_{i=0}^{N_1}\sum_{j=0}^{N_2} y^2(i, j)\hbar_1(i)\hbar_2(j) + \sum_{j=0}^{N_2}\sum_{i=0}^{N_1} h_1 \hbar_2(j) y^2_{\bar{x}_1}(i, j) + \\ + \sum_{i=0}^{N_1}\sum_{j=1}^{N_2} \hbar_1(i) h_2 y^2_{\bar{x}_2}(i, j)\Big]. \tag{27}$$

Montrons que

$$c_4 = \sqrt{2}(16 + l^2)/(l\sqrt{32 + l^2}), \quad l = \min(l_1, l_2). \tag{28}$$

En effet, de l'inégalité (36) du lemme 15, ch. V, pour $\varepsilon = \sqrt{2}$, on obtient

$$y^2(0, j) + y^2(N_1, j) \leqslant$$
$$\leqslant \frac{(16+l_1^2)\sqrt{2}}{l_1\sqrt{32+l_1^2}}\left[\sum_{i=1}^{N_1} h_1 y^2_{\bar{x}_1}(i, j) + \frac{1}{2}\sum_{i=0}^{N_1} \hbar_1(i)\, y^2(i, j)\right].$$

Notons que si l'on substitue ici l à l_1, l'inégalité ne fera que se renforcer. En multipliant maintenant le premier et le second membre de l'inégalité obtenue par $\hbar_2(j)$ et en sommant en j de 0 à N_2, on obtient

$$\sum_{j=0}^{N_2} \hbar_2(j)\,[y^2(0, j) + y^2(N_1, j)] \leqslant$$
$$\leqslant c_4\left[\sum_{j=0}^{N_2}\sum_{i=1}^{N_1} h_1\hbar_2(j)\, y^2_{\bar{x}_1}(i, j) + \frac{1}{2}\sum_{i=0}^{N_1}\sum_{j=0}^{N_2} y^2(i, j)\,\hbar_1(i)\,\hbar_2(j)\right], \qquad (29)$$

où c_4 est défini dans (28). De façon analogue, on trouve

$$\sum_{i=0}^{N_1} \hbar_1(i)\,[y^2(i, 0) + y^2(i, N_2)] \leqslant$$
$$\leqslant c_4\left[\sum_{i=0}^{N_1}\sum_{j=1}^{N_2} \hbar_1(i)\, h_2 y^2_{\bar{x}_2}(i, j) + \frac{1}{2}\sum_{i=0}^{N_1}\sum_{j=0}^{N_2} y^2(i, j)\,\hbar_1(i)\,\hbar_2(j)\right]. \qquad (30)$$

En additionnant (29) et (30), on obtient l'inégalité (27).

Pour résoudre l'équation (26), on peut recourir à la méthode itérative simple implicite (20), où $B = R$ et $\tau = 1/\gamma_2 = 1/(c_2(1 + c_4))$. Dans ce cas, en vertu du théorème 1, la méthode itérative (20) convergera dans H_B et, pour l'erreur, on aura l'estimation

$$\| y_n - u \|_B \leqslant \rho^n \| y_0 - u \|_B, \quad \rho = \sqrt{1-\xi}, \quad \xi = \gamma_1/\gamma_2 = c_1/(c_2(1 + c_4)).$$

Par conséquent, le nombre d'itérations $n_0(\varepsilon)$, exigé pour l'obtention de la précision relative ε, ne dépendra pas du nombre de nœuds dans le maillage $\bar{\omega}$.

Pour trouver y_{k+1}, on pose le problème

$$Ry_{k+1} = \varphi, \quad \varphi = Ry_k - \tau(Ay_k - f).$$

L'opérateur R correspondant au second problème aux limites pour une équation aux différences à coefficients constants, le problème mentionné peut être résolu par des méthodes directes décrites dans les chapitres III et IV en $O(N^2 \log_2 N)$ opérations arithmétiques ($N_1 = N_2 = N = 2^n$). Si les fonctions $k_\alpha(x, p)$ et $\varkappa_{\pm\alpha}(x, p_0)$ sont dérivables, l'opérateur A possède alors une dérivée Gâteau consti-

tuant un opérateur autoadjoint dans H au cas où sont remplies les conditions

$$a_{\alpha\beta}(x, p) = a_{\beta\alpha}(x, p), \quad \alpha, \beta = 0, 1, 2, \tag{31}$$

où $a_{\alpha\beta}(x, p) = \dfrac{\partial k_\alpha(x, p)}{\partial p_\beta}$. On peut montrer que si, outre (31), sont remplies les conditions

$$c_1 \sum_{\alpha=0}^{2} \xi_\alpha^2 \leqslant \sum_{\alpha, \beta=0}^{2} a_{\alpha\beta}(x, p)\,\xi_\alpha\xi_\beta \leqslant c_2 \sum_{\alpha=0}^{2} \xi_\alpha^2, \quad c_1 > 0,$$

$$0 \leqslant \frac{\partial \varkappa_{\pm\alpha}(x, p_0)}{\partial p_0} \leqslant c_3, \quad \alpha = 1, 2,$$

alors sont vérifiées les inégalités (15), où $\gamma_1 = c_1$, $\gamma_2 = c_2 + c_3c_4$, tandis que c_4 est défini dans (28). Dans ce cas, dans la méthode itérative (20) avec $B = R$, le paramètre τ peut être choisi égal à $\tau_0 = 2/(\gamma_1 + \gamma_2)$. En vertu du théorème 4 on aura pour l'erreur l'estimation

$$\| y_n - u \|_B \leqslant \rho_0^n \| y_0 - u \|_B, \quad \rho_0 = (1 - \xi)/(1 + \xi), \quad \xi = \gamma_1/\gamma_2.$$

Supposons qu'à présent il s'agit de trouver la solution du premier problème aux limites dans le rectangle $\bar{G}$

$$\sum_{\alpha=1}^{2} \frac{\partial}{\partial x_\alpha} k_\alpha\left(x, u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}\right) - k_0\left(x, u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}\right) = -\varphi(x), \quad x \in G,$$

$$u(x) = 0, \quad x \in \Gamma. \tag{32}$$

Posons que les fonctions $k_\alpha(x, p)$ sont continues en $p = (p_0, p_1, p_2)$ et que sont remplies les conditions

$$\sum_{\alpha=1}^{2} [k_\alpha(x, p) - k_\alpha(x, q)](p_\alpha - q_\alpha) \geqslant c_1 \sum_{\alpha=1}^{2} (p_\alpha - q_\alpha)^2, \quad c_1 > 0,$$

$$[k_0(x, p) - k_0(x, q)](p_0 - q_0) \geqslant 0, \tag{33}$$

$$\sum_{\alpha=0}^{2} [k_\alpha(x, p) - k_\alpha(x, q)]^2 \leqslant c_2 \sum_{\alpha=0}^{2} [k_\alpha(x, p) - k_\alpha(x, q)](p_\alpha - q_\alpha),$$

où $c_1 > 0$, $c_2 > 0$ pour $x \in \bar{G}$ et $|p|, |q| < \infty$.

Le problème (32) sur le maillage régulier rectangulaire $\bar{\omega} = \omega \cup \gamma$, introduit auparavant, sera mis en accord avec le schéma aux différences

$$\Lambda y = -f, \quad x \in \omega, \quad y(x) = 0, \quad x \in \gamma, \tag{34}$$

où $f = \varphi$, tandis que l'opérateur de différences Λ est défini de la façon suivante:

$$\Lambda y = \Lambda^- y = \frac{1}{2}\{[k_1(x, y, y_{\bar{x}_1}, y_{\bar{x}_2})]_{x_1} + [k_1(x, y, y_{x_1}, y_{x_2})]_{\bar{x}_1} + + [k_2(x, y, y_{\bar{x}_1}, y_{\bar{x}_2})]_{x_2} + [k_2(x, y, y_{x_1}, y_{x_2})]_{\bar{x}_2} - - k_0(x, y, y_{\bar{x}_1}, y_{\bar{x}_2}) - k_0(x, y, y_{x_1}, y_{x_2}]\}.$$

Fournissons encore deux approximations possibles:

$$\Lambda y = \Lambda^+ y = \frac{1}{2}\{[k_1(x, y, y_{\bar{x}_1}, y_{x_2})]_{x_1} + [k_1(x, y, y_{x_1}, y_{\bar{x}_2})]_{\bar{x}_1} + + [k_2(x, y, y_{x_1}, y_{\bar{x}_2})]_{x_2} + [k_2(x, y, y_{\bar{x}_1}, y_{x_2})]_{\bar{x}_2} - - k_0(x, y, y_{\bar{x}_1}, y_{x_2}) - k_0(x, y, y_{x_1}, y_{\bar{x}_2}]\}$$

et $\Lambda = \frac{1}{2}(\Lambda^- + \Lambda^+)$.

Dans l'exemple considéré H est l'espace des fonctions de mailles associées à ω et dont le produit scalaire se définit par la formule

$$(u, v) = \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} h_1 h_2 u(i, j) v(i, j).$$

Si dans les équations du schéma (34) on substitue $y|_\gamma = 0$, on obtient alors le schéma aux différences $\bar{\Lambda} y = -f$. En définissant l'opérateur A comme égal à $-\bar{\Lambda}$, on est en mesure d'écrire le schéma obtenu sous forme d'équation opératorielle (26) dans l'espace H.

En utilisant les conditions (33), on obtient pour les trois approximations que l'opérateur A satisfait aux inégalités (16), (17):

$$(Au - Av, u - v) \geqslant \gamma_1 (R(u - v), u - v), \quad \gamma_1 = c_1 > 0,$$
$$(R^{-1}(Au - Av), Au - Av) \leqslant \gamma_2 (Au - Av, u - v), \quad \gamma_2 = c_2(1 + c_4),$$

où

$$c_4 = \frac{1}{\delta}, \quad \delta = \frac{4}{h_1^2}\sin^2\frac{\pi h_1}{2l_1} + \frac{4}{h_2^2}\sin^2\frac{\pi h_2}{2l_2} \geqslant \frac{8}{l_1^2} + \frac{8}{l_2^2},$$

tandis que l'opérateur R correspond à l'opérateur de différences de Laplace $Ry = -\mathcal{R}\mathring{y}$, $y(x) = \mathring{y}(x)$ pour $x \in \omega$ et $\mathring{y}(x) = 0$ pour $x \in \gamma$, $\mathcal{R}u = u_{\bar{x}_1 x_1} + u_{\bar{x}_2 x_2}$.

Pour résoudre les équations (26), profitons de la méthode itérative simple (20) avec $B = R$ et $\tau = 1/\gamma_2$. En vertu du théorème 1 on aura l'estimation

$$\| y_n - u \|_B \leqslant \rho^n \| y_0 - u \|_B, \quad \rho = \sqrt{1 - \xi}, \quad \xi = \gamma_1/\gamma_2.$$

Comme auparavant, pour résoudre les équations $Ry_{k+1} = Ry_k - \tau (Ay_k - f)$, on peut utiliser les méthodes directes de réduction totale ou de séparation des variables proposées dans les chapitres III et IV.

4. Méthodes itératives pour des équations faiblement non linéaires. Dans le rectangle $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha, \ \alpha = 1, 2\}$ étudions l'équation elliptique faiblement non linéaire de second ordre

$$Lu = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} - k_0 \left(x, u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2} \right) = 0, \quad x \in G \tag{35}$$

aux conditions aux limites de première espèce

$$u(x) = 0, \quad x \in \Gamma. \tag{36}$$

La *faible non-linéarité* de l'équation (35) signifie que la fonction $k_0 (x, p_0, p_1, p_2)$ est définie pour $x \in \bar{G}$ et $|p_0|, |p_1|, |p_2| < \infty$, et est continue en x pour des p_0, p_1, p_2 fixés, et qu'il existe également des dérivées de la fonction $k_0 (x, p_0, p_1, p_2)$ en p_0, p_1 et p_2, qui satisfont aux conditions

$$c_2 \geqslant \frac{\partial k_0}{\partial p_0} \geqslant 0, \quad \left| \frac{\partial k_0}{\partial p_\alpha} \right| \leqslant M, \quad \alpha = 1, 2. \tag{37}$$

Sur le maillage $\bar{\omega} = \omega \cup \gamma$ régulier et rectangle, introduit auparavant, le schéma aux différences correspondant au problème (35), (36) prend la forme

$$\begin{gathered} \Lambda y = 0, \quad x \in \omega, \quad y(x) = 0, \quad x \in \gamma, \\ \Lambda y = \mathcal{R} y - \frac{1}{2} [k_0 (x, y, y_{\bar{x}_1}, y_{\bar{x}_2}) + k_0 (x, y, y_{x_1}, y_{x_2})], \end{gathered} \tag{38}$$

où $\mathcal{R} y = y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2}$ est l'opérateur de différences de Laplace.

Déterminons maintenant l'opérateur de différences $\Lambda' (v)$ dépendant de v:

$$\begin{aligned} \Lambda' (v) y = \mathcal{R} y - \frac{1}{2} [&a_{01} (x, v, v_{\bar{x}_1}, v_{\bar{x}_2}) y_{\bar{x}_1} + \\ &+ a_{01} (x, v, v_{x_1}, v_{x_2}) y_{x_1} + a_{02} (x, v, v_{\bar{x}_1}, v_{\bar{x}_2}) y_{\bar{x}_2} + \\ &+ a_{02} (x, v, v_{x_1}, v_{x_2}) y_{x_2} + (a_{00} (x, v, v_{\bar{x}_1}, v_{\bar{x}_2}) + a_{00} (x, v, v_{x_1}, v_{x_2})) y], \end{aligned}$$

où

$$a_{0\alpha} (x, p_0, p_1, p_2) = \frac{\partial k_0 (x, p_0, p_1, p_2)}{\partial p_\alpha}, \quad \alpha = 0, 1, 2.$$

Dans l'espace H des fonctions de mailles associées à ω définissons les opérateurs:

$$Ay = -\Lambda \overset{\circ}{y}, \quad Ry = -\mathcal{R} \overset{\circ}{y}, \quad A' (v) y = -\Lambda' (\overset{\circ}{v}) \overset{\circ}{y},$$

où

$$y(x) = \overset{\circ}{y}(x), \ v(x) = \overset{\circ}{v}(x) \text{ pour } x \in \omega$$

et

$$\overset{\circ}{y}(x) = 0, \quad \overset{\circ}{v}(x) = 0 \text{ pour } x \in \gamma.$$

L'opérateur $A'(v)$ est une dérivée Gâteau de l'opérateur A. En utilisant ces notations, écrivons le schéma aux différences sous forme de l'équation opératorielle (26).

Si $k_0(x, p_0, p_1, p_2)$ est indépendant de p_1, et p_2, c'est-à-dire si

$$k_0(x, p_0, p_1, p_2) = k_0(x, p_0),$$

alors on a

$$a_{01}(x, p) = a_{02}(x, p) = 0.$$

Dans ce cas l'opérateur $A'(v)$ est autoadjoint dans H.

En utilisant l'estimation inférieure de l'opérateur de différences $(-\mathcal{R})$

$$(-\mathcal{R}\overset{\circ}{y}, \overset{\circ}{y}) = -(\overset{\circ}{y}_{\bar{x}_1x_1} + \overset{\circ}{y}_{\bar{x}_2x_2}, \overset{\circ}{y}) \geqslant \delta(\overset{\circ}{y}, \overset{\circ}{y}),$$

où

$$\delta = \frac{4}{h_1^2}\sin^2\frac{\pi h_1}{2l_1} + \frac{4}{h_2^2}\sin^2\frac{\pi h_2}{2l_2} \geqslant \frac{8}{l_1^2} + \frac{8}{l_2^2},$$

les conditions (37) pour $M=0$ et les égalités

$$-(\Lambda'(\overset{\circ}{v})\overset{\circ}{y}, \overset{\circ}{y}) = -(\mathcal{R}\overset{\circ}{y}, \overset{\circ}{y}) + (a_{00}(x, \overset{\circ}{v})\overset{\circ}{y}, \overset{\circ}{y}),$$

on obtient

$$\gamma_1(Ry, y) \leqslant (A'(v)y, y) \leqslant \gamma_2(Ry, y),$$

où

$$\gamma_1 = 1, \quad \gamma_2 = 1 + c_2/\delta.$$

Par conséquent, si pour le cas considéré d'« autoconjugaison » on utilise la méthode itérative (20) avec $B = D = R$ et $\tau = \tau_0 = 2/(\gamma_1 + \gamma_2)$, alors, en vertu du théorème 2, on aura pour l'erreur l'estimation

$$\| y_n - u \|_B \leqslant \rho_0^n \| y_0 - u \|_B, \quad \rho_0 = (1-\xi)/(1+\xi), \quad \xi = \gamma_1/\gamma_2.$$

L'opérateur R dans le schéma (20) peut être inverti au moyen de l'une des méthodes directes.

36—01162

CHAPITRE XIV

EXEMPLES DE RÉSOLUTION DES ÉQUATIONS ELLIPTIQUES DE MAILLES

On étudie dans le § 1 quelques procédés de construction des schémas itératifs implicites, en particulier, avec le recours à un régularisateur. Le § 2 est consacré à l'étude des méthodes de résolution des systèmes d'équations elliptiques. On y examine comment la théorie générale s'applique à la résolution de quelques problèmes de la théorie de l'élasticité.

§ 1. Procédés de construction des schémas itératifs implicites

1. Principe de régularisation dans la théorie générale des méthodes itératives. On a exposé dans les chapitres VI—VIII, XII, XIII la théorie générale des méthodes itératives utilisées pour la résolution de l'équation opératorielle

$$Au = f. \tag{1}$$

Dans la théorie générale des méthodes itératives on n'a pas utilisé la structure concrète des opérateurs du schéma itératif, la théorie ne recourant qu'à un minimum d'information de nature fonctionnelle générale sur les opérateurs. Cela permet (une fois fixés les opérateurs du schéma) d'indiquer les principes généraux de construction des méthodes itératives optimales. Par exemple, si les opérateurs A et B du schéma itératif à deux couches

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}}+Ay_k=f, \quad k=0, 1, \ldots, \quad y_0\in H \tag{2}$$

satisfont aux conditions

$$B = B^* > 0, \quad A = A^* > 0, \tag{3}$$

$$\gamma_1 B\leqslant A\leqslant \gamma_2 B, \quad \gamma_1 > 0, \tag{4}$$

le jeu des paramètres d'itération de Tchébychev τ_k:

$$\tau_k=\frac{\tau_0}{1+\rho_0\mu_k}, \quad \mu_k\in \mathfrak{M}_n=\left\{-\cos\frac{(2i-1)\pi}{2n}, \quad 1\leqslant i\leqslant n\right\}, \ 1\leqslant k\leqslant n,$$

où

$$\tau_0=\frac{2}{\gamma_1+\gamma_2}, \quad \rho_0=\frac{1-\xi}{1+\xi}, \quad \xi=\frac{\gamma_1}{\gamma_2}$$

est alors le meilleur.

A quelles exigences doit-on se conformer lors du choix de l'opérateur B ? On a remarqué dans le § 3, ch. V, que le choix de B doit se plier à deux exigences : 1) garantir la convergence la plus rapide de la méthode ; 2) veiller à ce que l'inversion de cet opérateur soit économique.

Pour l'exemple donné plus haut, la première exigence est satisfaite si l'énergie de l'opérateur B est proche de celle de l'opérateur A, c'est-à-dire si dans les inégalités (4) γ_1 et γ_2 sont proches. Pour remplir la seconde exigence, il faut de la classe des opérateurs B, proches quant à leur énergie de l'opérateur A, choisir celui dont l'inversion est la plus facile.

Comment construire les opérateurs facilement inversibles? Il est évident que si $B^1, B^2, \ldots, B^p$ sont des opérateurs facilement inversibles, l'opérateur $B = B^1B^2 \ldots B^p$, constituant leur produit, est également facilement inversible.

Notons qu'à la différence des facteurs l'opérateur B lui-même peut posséder une structure complexe. Par exemple, soit $B^\alpha = E + \omega R_\alpha$, $\alpha = 1, 2$, où R_α est un opérateur correspondant à l'opérateur de différences $(-\mathscr{R}_\alpha)$: $\mathscr{R}_\alpha\ y = y_{\bar{x}_\alpha x_\alpha}$, $\alpha = 1, 2$. A l'opérateur B^α correspond un opérateur de différences triponctuel qui s'inverse par la méthode du balayage en un nombre d'opérations arithmétiques proportionnel à celui d'inconnues dans le problème. L'opérateur $B = B^1B^2$ possède un stencil à neuf points et il lui correspond l'opérateur de différences $\mathscr{B}$:

$$\mathscr{B}y = y - \omega \sum_{\alpha=1}^{2} y_{\bar{x}_\alpha x_\alpha} + \omega^2 y_{\bar{x}_1 x_1 \bar{x}_2 x_2}.$$

La complication de la structure de l'opérateur B permet d'accoroître le rapport $\xi = \gamma_1/\gamma_2$ et d'augmenter ainsi la vitesse de convergence de la méthode itérative.

Pour construire l'opérateur B on peut partir d'un opérateur quelconque $R = R^* > 0$ (d'un régularisateur) qui est énergétiquement équivalent à A et B:

$$c_1 R \leqslant A \leqslant c_2 R, \quad c_2 \geqslant c_1 > 0, \tag{5}$$

$$\mathring{\gamma}_1 B \leqslant R \leqslant \mathring{\gamma}_2 B, \quad \mathring{\gamma}_2 \geqslant \mathring{\gamma}_1 > 0. \tag{6}$$

Dans ce cas les inégalités (4) avec les constantes $\gamma_1 = c_1\mathring{\gamma}_1$, $\gamma_2 = c_2\mathring{\gamma}_2$ se vérifient avec

$$\xi = \gamma_1/\gamma_2 = (c_1/c_2)\,\mathring{\xi}, \quad \mathring{\xi} = \mathring{\gamma}_1/\mathring{\gamma}_2.$$

En quoi consiste l'idée de l'introduction du régularisateur R ? Généralement, pour les problèmes aux limites elliptiques associés à un maillage, l'opérateur R est choisi de manière que les constantes c_1 et c_2 des inégalités (5) soient indépendantes des paramètres du maillage (du nombre des nœuds du maillage). Par exemple, si l'opéra-

teur A correspond à un opérateur de différences à coefficients variables

$$\Lambda y = (a_1 y_{\bar{x}_1})_{x_1} + (a_2 y_{\bar{x}_2})_{x_2}, \quad 0 < c_1 \leqslant a_\alpha \leqslant c_2,$$

associé à un maillage régulier $\bar{\omega} = \{x_{ij} = (ih_1, jh_2), 0 \leqslant i \leqslant N_1, 0 \leqslant j \leqslant N_2, h_\alpha N_\alpha = l_\alpha, \alpha = 1, 2\}$ introduit dans le rectangle $\bar{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha, \alpha = 1, 2\}$, de sorte que $Ay = -\Lambda \mathring{y}$, où $y(x) = \mathring{y}(x)$ pour $x \in \omega$ et $\mathring{y}(x) = 0$ pour $x \in \gamma$, on peut choisir en guise de R l'opérateur correspondant à l'opérateur de différences de Laplace $\mathcal{R} y = (\mathcal{R}_1 + \mathcal{R}_2) y = y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2}$, $Ry = -\mathcal{R}\mathring{y}$, où les opérateurs $\mathcal{R}_\alpha$ sont définis plus haut.

En utilisant les formules de différences de Green, on montre sans peine (voir point 8, § 2, ch. V) que les opérateurs A et B sont autoadjoints dans H et que les inégalités (5) sont satisfaites. H est ici l'espace des fonctions de mailles associées à ω et dont le produit scalaire se détermine par la formule $(u, v) = \sum_{x \in \omega} u(x) v(x) h_1 h_2$.

Supposons à présent que l'opérateur A correspond à l'opérateur de différences elliptique contenant des dérivées mixtes

$$\Lambda y = \sum_{\alpha, \beta = 1}^{2} 0{,}5 \left[(k_{\alpha\beta} y_{\bar{x}_\beta})_{x_\alpha} + (k_{\alpha\beta} y_{x_\beta})_{\bar{x}_\alpha} \right],$$

et que sont remplies les conditions de forte ellipticité:

$$c_1 \sum_{\alpha=1}^{2} \xi_\alpha^2 \leqslant \sum_{\alpha, \beta = 1}^{2} k_{\alpha\beta}(x) \xi_\alpha \xi_\beta \leqslant c_2 \sum_{\alpha=1}^{2} \xi_\alpha^2, \quad c_1 > 0.$$

Prenons en guise de régularisateur l'opérateur R défini plus haut. Au point 8, § 2, ch. V on a montré que pour les opérateurs A et R examinés les inégalités (5) sont satisfaites.

Donnons encore un exemple. Supposons que l'opérateur A correspond à l'opérateur de différences de Laplace d'ordre de précision élevé

$$\Lambda y = x_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} + \frac{h_1^2 + h_2^2}{12} y_{\bar{x}_1 x_1 \bar{x}_2 x_2}.$$

Montrons que si en guise d'opérateur R on choisit l'opérateur mentionné plus haut, les inégalités (5) à constantes $c_1 = 2/3$, $c_2 = 1$ se vérifient.

En effet, en utilisant la première formule de différences de Green et l'égalité $y_{\bar{x}_1 x_1 \bar{x}_2 x_2} = y_{\bar{x}_1 \bar{x}_2 x_1 x_2}$, qui se vérifie pour les fonctions de mailles associées à un maillage rectangulaire $\bar{\omega}$, on aboutit à

$$\begin{aligned} -(\Lambda \mathring{y}, \mathring{y}) &= (\mathring{y}_{\bar{x}_1}^2, 1)_1 + (\mathring{y}_{\bar{x}_2}^2, 1)_2 - \frac{h_1^2 + h_2^2}{12} (\mathring{y}_{\bar{x}_1 \bar{x}_2}^2, 1)_{12}, \\ -(\mathcal{R} \mathring{y}, \mathring{y}) &= (\mathring{y}_{\bar{x}_1}^2, 1)_1 + (\mathring{y}_{\bar{x}_2}^2, 1)_2. \end{aligned} \tag{7}$$

On a adopté ici comme notations:

$$(u, v)_1 = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2-1} u(i, j) v(i, j) h_1 h_2,$$

$$(u, v)_2 = \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2} u(i, j) v(i, j) h_1 h_2,$$

$$(u, v)_{12} = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} u(i, j) v(i, j) h_1 h_2.$$

A partir de (7) on tire l'estimation $A \leqslant R$, c'est-à-dire que dans (5) $c_2 = 1$. Ensuite, compte tenu de $\overset{\circ}{y}_{\bar{x}_2}(x) = 0$ pour $x_1 = 0$, l_1 et $\overset{\circ}{y}_{\bar{x}_1} = 0$ pour $x_2 = 0$, l_2, du lemme 12, ch. V, on obtient l'estimation

$$(\overset{\circ}{y}{}^2_{\bar{x}_1\bar{x}_2}, 1)_{12} \leqslant \frac{4}{h_2^2} (\overset{\circ}{y}{}^2_{\bar{x}_1}, 1)_1 \tag{8}$$

et de façon analogue

$$(\overset{\circ}{y}{}^2_{\bar{x}_1\bar{x}_2}, 1)_{12} = (\overset{\circ}{y}{}^2_{\bar{x}_2\bar{x}_1}, 1)_{12} \leqslant \frac{4}{h_1^2} (\overset{\circ}{y}{}^2_{\bar{x}_2}, 1)_2. \tag{9}$$

En multipliant (8) par $h_2^2/12$ et (9) par $h_1^2/12$, puis, en additionnant les inégalités ainsi obtenues, il vient

$$\frac{h_1^2 + h_2^2}{12} (\overset{\circ}{y}{}^2_{\bar{x}_1\bar{x}_2}, 1)_{12} \leqslant \frac{1}{3} [(\overset{\circ}{y}{}^2_{\bar{x}_1}, 1)_1 + (\overset{\circ}{y}{}^2_{\bar{x}_2}, 1)_2].$$

De là et à partir de (7) on déduit l'estimation $A \geqslant 2R/3$. La proposition est démontrée.

Les exemples examinés montrent qu'on peut choisir en guise de régularisateur pour des opérateurs A différents le même opérateur R. Aussi le problème de construction de l'opérateur B se simplifie-t-il pour le schéma itératif implicite. L'opérateur B se construit sur la base de sa proximité en énergie du régularisateur R. La classe des régularisateurs est essentiellement plus étroite que la classe comprenant les opérateurs A. Si l'opérateur B est choisi et, par suite, les constantes $\overset{\circ}{\gamma}_1$ et $\overset{\circ}{\gamma}_2$ des inégalités (6) sont trouvées, il ne reste qu'à obtenir pour chaque opérateur concret A les constantes c_1 et c_2 dans les inégalités (5).

Avec l'utilisation du régularisateur, la difficulté principale consiste dans l'obtention des estimations pour $\overset{\circ}{\gamma}_1$ et $\overset{\circ}{\gamma}_2$. Le plus souvent l'opérateur B prend une forme factorisée, les facteurs dépendant de certains paramètres d'itération. On définit ainsi une famille d'opérateurs B de structure déferminée et caractérisée par les paramètres mentionnés. Ces paramètres doivent être choisis sur la base de la condition de maximum de $\overset{\circ}{\xi}$. On étudiera quelques exemples d'opérateurs factorisés au point suivant. En attendant, notons qu'en guise d'opérateur B on peut quelquefois choisir le régularisateur R ($\overset{\circ}{\gamma}_1 = \overset{\circ}{\gamma}_2 = 1$).

2. Schémas itératifs à opérateur factorisé. Au point 1 le principe de régularisation a été illustré par un exemple d'opérateur A autoadjoint. Dans ce cas les inégalités (4), s'ensuivent des inégalités (5) et (6).

On a montré dans le ch. VI que si l'opérateur A n'est pas autoadjoint dans H, tandis que l'espace énergétique H_D est engendré par un opérateur D autoadjoint et défini positif, où D est soit B, soit $A^*B^{-1}A$, il est nécessaire de substituer aux inégalités (4) les inégalités

$$\gamma_1(Bx, x) \leqslant (Ax, x), \quad (B^{-1}Ax, Ax) \leqslant \gamma_2(Ax, x), \quad \gamma_1 > 0 \quad (10)$$

ou bien les inégalités

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \quad (B^{-1}A_1x, A_1x) \leqslant \gamma_3^2 (Bx, x), \quad \gamma_1 > 0, \quad (11)$$

où $A_1 = 0,5\,(A - A^*)$ est la partie non adjointe de l'opérateur A. Supposons que l'opérateur $B = B^* > 0$ est construit sur la base du régularisateur R et que les inégalités (6) sont satisfaites. Alors, si l'opérateur R satisfait aux conditions

$$c_1 (Rx, x) \leqslant (Ax, x), \ (R^{-1}Ax, Ax) \leqslant c_2 (Ax, x), \ c_1 > 0. \quad (10')$$

on obtient alors les inégalités (10) aux constantes $\gamma_1 = c_1\mathring{\gamma}_1$, $\gamma_2 = c_2\mathring{\gamma}_2$.

En effet, à partir du lemme 9, ch. V, et de l'inégalité (6) il résulte que les inégalités $\mathring{\gamma}_1 R^{-1} \leqslant B^{-1} \leqslant \mathring{\gamma}_2 \cdot R^{-1}$ sont satisfaites. De là il vient

$$(B^{-1}Ax, Ax) \leqslant \mathring{\gamma}_2\ (R^{-1}Ax, Ax) \leqslant c_2\mathring{\gamma}_2\ (Ax, x).$$

De façon analogue on démontre que si l'opérateur R vérifie les conditions

$$c_1 R \leqslant A \leqslant c_2 R, \quad (R^{-1}A_1x, A_1x) \leqslant c_3^2 (Rx, x), \quad c_1 > 0, \quad (11')$$

les inégalités (11) aux constantes $\gamma_1 = c_1\mathring{\gamma}_1$, $\gamma_2 = c_2\mathring{\gamma}_2$, $\gamma_3 = c_3\mathring{\gamma}_2$ sont également satisfaites.

Ainsi donc, dans le cas d'un opérateur A non autoadjoint également il faut savoir obtenir les estimations pour $\mathring{\gamma}_1$ et $\mathring{\gamma}_2$ figurant dans les inégalités (6).

Essayons maintenant d'obtenir les inégalités (6) pour les opérateurs autoadjoints R et B. Examinons deux cas:

1) L'opérateur R se présente sous forme d'une somme $R = R_1 + R_2$ d'opérateurs R_1 et R_2 mutuellement autoadjoints:

$$R_2 = R_1^*, \quad (12)$$

de sorte que $(R_1x, x) = (R_2x, x) = 0,5\,(Rx, x)$, $x \in H$, tandis que l'opérateur B est de l'aspect

$$B = (E + \omega R_1)(E + \omega R_2), \quad (13)$$

où $\omega > 0$ est un paramètre.

2) L'opérateur R se présente sous forme de la somme $R = R_1 + R_2 + \ldots + R_p$, $p \geqslant 2$, d'opérateurs autoadjoints deux à deux permutables R_α, $\alpha = 1, 2, \ldots, p$, de sorte que

$$R_\alpha = R_\alpha^*, \quad R_\alpha R_\beta = R_\beta R_\alpha, \quad \alpha, \beta = 1, 2, \ldots, p, \tag{14}$$

quant à l'opérateur B, il est factorisé et prend la forme

$$B = \prod_{\alpha=1}^{p} (E + \omega R_\alpha), \tag{15}$$

où $\omega > 0$ est un paramètre.

Dans chaque cas l'opérateur B est autoadjoint dans H. Soulignons spécialement le caractère universel du choix de l'opérateur B en la forme (13), où les opérateurs R_1 et R_2 satisfont à la condition (12).

La question est d'obtenir les estimations pour $\overset{\circ}{\gamma}_1$ et $\overset{\circ}{\gamma}_2$ figurant dans (6), ainsi que de choisir le paramètre d'itération ω sur la base de la condition du maximum du rapport $\overset{\circ}{\xi} = \overset{\circ}{\gamma}_1/\overset{\circ}{\gamma}_2$.

Etudions séparément chaque cas. Le premier cas a été l'objet d'une étude détaillée au ch. X consacré à la méthode triangulaire alternée. On se limitera donc ici à la formulation des résultats.

Théorème 1. *Supposons que les conditions* (12) *sont remplies et que dans les inégalités*

$$R \geqslant \delta E, \quad (R_2 x, R_2 x) \leqslant \frac{\Delta}{4} (Rx, x), \quad \delta > 0 \tag{16}$$

les constantes δ *et* Δ *sont données. Dans ce cas, pour la valeur optimale du paramètre* $\omega = \omega_0 = 2/\sqrt{\delta\Delta}$, *l'opérateur* B, *défini par l'égalité* (13), *satisfait aux inégalités* (6) *avec les constantes*

$$\overset{\circ}{\gamma}_1 = \frac{\delta}{2(1+\sqrt{\eta})}, \quad \overset{\circ}{\gamma}_2 = \frac{\delta}{4\sqrt{\eta}}, \quad \eta = \frac{\delta}{\Delta}.$$

Notons qu'on peut procéder à l'examen de la forme de l'opérateur B plus générale que (13), à savoir:

$$B = (\mathscr{D} + \omega R_1)\mathscr{D}^{-1}(\mathscr{D} + \omega R_2),$$

où $\mathscr{D} = \mathscr{D}^* > 0$. Il s'ensuit du lemme 1, ch. X, que le théorème 1 reste vrai, il ne faut que remplacer (16) par les inégalités suivantes:

$$R \geqslant \delta \mathscr{D}, \quad (\mathscr{D}^{-1} R_2 x, R_2 x) \leqslant \frac{\Delta}{4} (Rx, x).$$

$\mathscr{D}$ y joue le rôle d'un paramètre d'itération auxiliaire.

L'opérateur B s'inverse facilement dans le cas, par exemple, où à l'opérateur R_1 correspond une matrice triangulaire inférieure, à R_2 une matrice triangulaire supérieure, et à $\mathscr{D}$ une matrice diago-

nale. Si l'opérateur R correspond à un opérateur de différences elliptique, les matrices triangulaires mentionnées posséderont sur chaque ligne un nombre fini d'éléments non nuls, indépendant de celui des nœuds dans le maillage. L'inversion de chaque facteur figurant dans l'opérateur B peut donc se réaliser en un nombre d'opérations proportionnel à celui d'inconnues du problème.

Passons à présent au second cas.

Théorème 2. *Soient l'opérateur B présenté sous la forme* (15), *les conditions* (14) *remplies et les bornes des opérateurs R_α données:*

$$\delta_\alpha E \leqslant R_\alpha \leqslant \Delta_\alpha E, \quad \delta_\alpha > 0, \quad \alpha = 1, 2, \ldots, p.$$

Dans ce cas, pour la valeur optimale du paramètre ω

$$\omega = \omega_0 = \frac{1}{\Delta} \frac{1-\eta^{1/p}}{\eta^{1/p}-\eta},$$

l'opérateur B vérifie les inégalités (6) *aux constantes*

$$\overset{\circ}{\gamma}_1 = \frac{p\Delta}{(1+\omega_0\Delta)^p}, \quad \overset{\circ}{\gamma}_2 = \overset{\circ}{\gamma}_1 \frac{p-k(1-\eta)}{p\eta^{k/p}}, \quad \eta = \frac{\delta}{\Delta},$$

où

$$\delta = \min_\alpha \delta_\alpha, \quad \Delta = \max_\alpha \Delta_\alpha, \quad k = \left[\frac{p}{1-\eta} - \frac{\eta^{1/p}}{1-\eta^{1/p}}\right],$$

$[a]$ étant une partie entière du nombre a.

La démonstration étant laborieuse, on s'abstient de la donner. Notons seulement qu'en vertu des conditions (14) l'opérateur B est permutable avec les opérateurs R_α, $\alpha = 1, 2, \ldots, p$, et donc,

$$\overset{\circ}{\gamma}_1 = \min_{\delta \leqslant x_\alpha \leqslant \Delta} \frac{x_1+x_2+\ldots+x_p}{\prod_{\alpha=1}^{p}(1+\omega x_\alpha)}, \quad \overset{\circ}{\gamma}_2 = \max_{\delta \leqslant x_\alpha \leqslant \Delta} \frac{x_1+x_2+\ldots+x_p}{\prod_{\alpha=1}^{p}(1+\omega x_\alpha)}.$$

Notons les cas particuliers du théorème 2. Si $p=2$, alors

$$k=1, \quad \omega_0 = \frac{1}{\sqrt{\delta\Delta}}, \quad \overset{\circ}{\gamma}_1 = \frac{2\delta}{(1+\sqrt{\eta})^2}, \quad \overset{\circ}{\gamma}_2 = \frac{\delta}{\sqrt{\eta}} \frac{1+\eta}{(1+\sqrt{\eta})^2}.$$

Si $p=3$, on a alors

$$k=2, \quad \omega_0 = \frac{1}{\sqrt[3]{\delta\Delta}\,(\sqrt[3]{\delta}+\sqrt[3]{\Delta})}, \quad \overset{\circ}{\gamma}_1 = 3\delta\left(\frac{1-\eta^{2/3}}{1-\eta}\right)^3,$$

$$\overset{\circ}{\gamma}_2 = \frac{\delta(1+2\eta)}{\eta^{2/3}}\left(\frac{1-\eta^{2/3}}{1-\eta}\right)^3.$$

Pour le cas où $p = 2$, on peut obtenir des meilleures estimations pour $\overset{\circ}{\gamma}_1$ et $\overset{\circ}{\gamma}_2$ en introduisant dans l'opérateur

$$B = (E + \omega_1 R_1)(E + \omega_2 R_2) \tag{17}$$

deux paramètres ω_1 et ω_2 qui prennent en compte le fait que les bornes des opérateurs R_1 et R_2 sont différentes. On a ainsi le théorème 3.

Théorème 3. *Supposons que l'opérateur B est de la forme* (17), *les conditions*

$$R_\alpha = R_\alpha^*, \quad \alpha = 1, 2, \quad R_1R_2 = R_2R_1$$

sont satisfaites et les bornes des opérateurs R_1 et R_2 données:

$$\delta_\alpha E \leqslant R_\alpha \leqslant \Delta_\alpha E, \quad \alpha = 1, 2, \quad \delta_1 + \delta_2 > 0.$$

Dans ce cas, pour des valeurs optimales des paramètres ω_1 et ω_2

$$\omega_1 = \frac{1+t\sqrt{\eta}}{r\sqrt{\eta}+s}, \quad \omega_2 = \frac{1-t\sqrt{\eta}}{r\sqrt{\eta}-s},$$

les inégalités (6) *sont satisfaites et possèdent les constantes*

$$\mathring{\gamma}_1 = \frac{4\sqrt{\eta}}{(\omega_1+\omega_2)(1+\sqrt{\eta})^2}, \quad \mathring{\gamma}_2 = \frac{2(1+\eta)}{(\omega_1+\omega_2)(1+\sqrt{\eta})^2},$$

où

$$r = \frac{\Delta_2+\Delta_1 b}{1+b}, \quad s = \frac{\Delta_2-\Delta_1 b}{1+b}, \quad t = \frac{1-b}{1+b}, \quad \eta = \frac{1-a}{1+a},$$

$$a = \sqrt{\frac{(\Delta_1-\delta_1)(\Delta_2-\delta_2)}{(\Delta_1+\delta_2)(\Delta_2+\delta_1)}}, \quad b = \frac{\Delta_2+\delta_1}{\Delta_1-\delta_1}\, a.$$

Pour démontrer le théorème, effectuons la substitution en posant

$$R_1 = (r\bar{R}_1 - sE)(E - t\bar{R}_1)^{-1}, \quad R_2 = (r\bar{R}_2 + sE)(E + t\bar{R}_2)^{-1}$$

où r, s, t ont les valeurs indiquées. On peut montrer que les opérateurs ainsi définis:

$$\bar{R}_1 = (R_1 + sE)(rE + tR_1)^{-1}, \quad \bar{R}_2 = (R_2 - sE)(rE - tR_2)^{-1}$$

vérifient les conditions $\bar{R}_\alpha = \bar{R}_\alpha^*$, $\alpha = 1, 2$, $\bar{R}_1\bar{R}_2 = \bar{R}_2\bar{R}_1$ et possèdent les mêmes bornes $\eta E \leqslant \bar{R}_\alpha \leqslant E$, $\eta > 0$, $\alpha = 1, 2$. Ensuite, vu que les opérateurs $E - t\bar{R}_1$ et $E + t\bar{R}_2$ sont autoadjoints et définis positifs, il existe des opérateurs permutables tels que $(E - t\bar{R}_1)^{1/2}$ et $(E + t\bar{R}_2)^{1/2}$. Posons

$$x = (E - t\bar{R}_1)^{1/2}(E + t\bar{R}_2)^{1/2}y.$$

On obtient

$$(Bx, x) = (1 - \omega_1 s)(1 + \omega_2 s)(\bar{B}y, y), \tag{18}$$

$$(Rx, x) = (r - st)(\bar{R}y, y), \tag{19}$$

où $\bar{B} = (E + \bar{\omega}\bar{R}_1)(E + \bar{\omega}\bar{R}_2)$, $\bar{R} = \bar{R}_1 + \bar{R}_2$,

$$\bar{\omega} = \frac{\omega_1 r - t}{1-\omega_1 s} = \frac{\omega_2 r + t}{1+\omega_2 s}. \tag{20}$$

A partir de (20) on trouve

$$2\bar{\omega} = \frac{\omega_1 r - t}{1-\omega_1 s} + \frac{\omega_2 r + t}{1+\omega_2 s} = \frac{(r-st)(\omega_1+\omega_2)}{(1-\omega_1 s)(1+\omega_2 s)}.$$

De là et à partir de (18), (19), il vient

$$\frac{(Rx, x)}{(Bx, x)} = \frac{2\bar{\omega}}{\omega_1 + \omega_2} \frac{(\bar{R}y, y)}{(\bar{B}y, y)}. \tag{21}$$

En utilisant le théorème 2, on obtient que pour

$$\bar{\omega} = \omega_0 = 1/\sqrt{\eta} \tag{22}$$

on a les inégalités

$$\mathring{\bar{\gamma}}_1 (\bar{R}y, y) \leqslant (\bar{B}y, y) \leqslant \mathring{\bar{\gamma}}_2 (\bar{R}y, y), \tag{23}$$

où

$$\mathring{\bar{\gamma}}_1 = \frac{2\eta}{(1+\sqrt{\eta})^2}, \quad \mathring{\bar{\gamma}}_2 = \frac{\sqrt{\eta}(1+\eta)}{(1+\sqrt{\eta})^2}.$$

Par conséquent, à partir de (20) et (22) on déduit les valeurs optimales des paramètres ω_1 et ω_2:

$$\omega_1 = \frac{1+t\sqrt{\eta}}{r\sqrt{\eta}+s}, \quad \omega_2 = \frac{1-t\sqrt{\eta}}{r\sqrt{\eta}-s},$$

tandis que de (21) et (23) s'ensuivent les inégalités (6) aux constantes $\mathring{\gamma}_1$ et $\mathring{\gamma}_2$ indiquées lors de l'énoncé du théorème 3. Le théorème 3 est démontré.

3. Procédé d'inversion implicite de l'opérateur B (méthode à deux étapes). On a étudié au point 2 le mode de construction des schémas itératifs implicites, qui se caractérise par le fait que l'opérateur B est donné de façon constructive sous la forme d'un produit d'opérateurs facilement inversibles. Examinons encore un procédé, avec lequel l'approximation itérative y_{k+1} s'obtient au moyen d'une procédure auxiliaire qui peut être assimilée à une inversion implicite d'un certain opérateur B.

Rappelons que l'idée générale de ce procédé a été étudiée au point 4, § 3, ch. V. Au point 4, § 1, ch. XIII ce procédé a été appliqué à la construction de la méthode itérative de résolution de l'équation à opérateur A non linéaire. On y a également formulé les conditions permettant d'obtenir les estimations de $\mathring{\gamma}_1$ et $\mathring{\gamma}_2$ entrant dans les inégalités (6).

Exposons les résultats obtenus. Supposons que l'approximation itérative y_{k+1} est obtenue suivant la formule du schéma avec correction $y_{k+1} = y_k - \tau_{k+1} w^p$, la correction w^p étant la solution approchée de l'équation auxiliaire

$$Rw = r_k, \quad r_k = Ay_k - f. \tag{24}$$

R est ici le régularisateur satisfaisant aux inégalités (5) au cas d'un opérateur A autoadjoint et vérifiant les inégalités (10′) ou (11′) pour un opérateur A non autoadjoint.

Supposons que l'équation (24) se résout à l'aide d'un schéma itératif à deux couches, de sorte que l'erreur $z^m = w^m - w$ vérifie l'équation

$$z^{m+1} = S_{m+1} z^m, \quad m = 0, 1, \ldots, p-1, \quad z^0 = w^0 - w,$$

où S_{m+1} est l'opérateur de passage de la m-ième à la $(m+1)$-ième itération.

En choisissant $w^0 = 0$, il résulte des égalités

$$z^p = w^p - w = T_p (w^0 - w), \quad T_p = \prod_{m=1}^{p} S_m, \quad w = R^{-1} r_k,$$

que

$$w^p = B^{-1} r_k, \text{ où } B = R (E - T_p)^{-1}.$$

En portant l'expression trouvée pour w^p dans (23), on aboutit au schéma itératif implicite (2) avec opérateur mentionné B.

Théorème 4. *Soient remplies les conditions*

$$R = R^* > 0, \quad T_p^* R = R T_p, \quad \| T_p \|_R \leqslant q < 1.$$

Alors l'opérateur $B = R (E - T_p)^{-1}$ est un opérateur autoadjoint et défini positif dans H et les inégalités (6) *avec les constantes $\overset{\circ}{\gamma}_1 = 1 - q$, $\overset{\circ}{\gamma}_2 = 1 + q$ sont satisfaites.*

Pour esquisser la démonstration, voir lemme 2, ch. XIII.

Remarque. Si les opérateurs R et T_p sont autoadjoints et permutables et $\| T_p \| \leqslant q < 1$, les assertions du théorème 4 sont alors vraies.

Par le procédé décrit plus haut, on a ainsi construit le schéma itératif implicite à deux couches. Mais si l'on part des formules

$$y_{k+1} = \alpha_{k+1} y_k + (1 - \alpha_{k+1}) y_{k-1} - \tau_{k+1} \alpha_{k+1} w_k^p, \quad k = 1, 2, \ldots,$$
$$y_1 = y_0 - \tau_1 w_0^p,$$

et que l'on obtienne l'erreur w_k^p pour tout $k = 0, 1, \ldots$ comme une solution approchée de l'équation (24), on aboutit au schéma itératif implicite à trois couches

$$B y_{k+1} = \alpha_{k+1} (B - \tau_{k+1} A) y_k + (1 - \alpha_{k+1}) B y_{k-1} + \alpha_{k+1} \tau_{k+1} f,$$
$$B y_1 = (B - \tau_1 A) y_0 + \tau_1 f. \tag{25}$$

Notons en conclusion que les paramètres d'itération τ_k du schéma (2) et τ_k, α_k du schéma (25) sont choisis en conformité avec la théorie générale des méthodes itératives. Il se pose alors le problème de choix du nombre optimal d'itérations p pour le processus itératif

auxiliaire. Eclairons la situation. Pour simplifier, on admet que le processus auxiliaire est stationnaire ($S_m \equiv S$), les opérateurs R et S sont autoadjoints et permutables et la condition $\| S \| \leqslant \rho$ est vérifiée. Alors $q = \rho^p$, c'est-à-dire

$$p = \ln q / \ln \rho. \tag{26}$$

Les opérateurs A et B vérifient les inégalités (4) aux constantes

$$\gamma_1 = c_1 (1 - q), \quad \gamma_2 = c_2 (1 + q).$$

Si les paramètres d'itération τ_k du schéma (2) sont choisis suivant les formules de la méthode de Tchébychev, on a alors pour le nombre d'itérations l'estimation

$$n \geqslant n_0 (\varepsilon), \; n_0 (\varepsilon) = \ln (0{,}5 \, \varepsilon) / \ln \rho_1,$$

où $\rho_1 = \rho_1 (q) = \dfrac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}$, $\quad \xi = \dfrac{\gamma_1}{\gamma_2} = \dfrac{c_1}{c_2} \dfrac{1-q}{1+q}$. Dans ce cas le nombre total d'itérations $k = pn$ est estimé à

$$k \geqslant k_0 (\varepsilon), \quad k_0 (\varepsilon) = \frac{\ln 0{,}5\varepsilon}{\ln \rho} \, \frac{\ln q}{\ln \rho_1 (q)}.$$

Il en résulte que la quantité q définissant, suivant (26), le nombre d'itérations internes doit être choisie sur la base de la condition du minimum de la fonction $\varphi (q) = \ln q / \ln \rho_1 (q)$. Ce problème peut être résolu numériquement.

§ 2. Systèmes d'équations elliptiques

1. Problème de Dirichlet pour un système d'équations elliptiques dans un parallélépipède à p dimensions. Soient $u = (u^1(x), u^2 (x), \ldots, u^{m_0} (x))$ et $f = (f^1 (x), f^2 (x), \ldots, f^{m_0} (x))$ des vecteurs de dimension m_0, $x = (x_1, x_2, \ldots, x_p)$ le point d'un espace de dimension p, $k = (k_{\alpha\beta})$ la matrice maillée de dimension $p \times p$, de sorte que la maille $k_{\alpha\beta} = (k_{\alpha\beta}^{sm} (x))$ constitue une matrice de dimension $m_0 \times m_0$:

$$k = \begin{Vmatrix} k_{11} & k_{12} & \ldots & k_{1p} \\ k_{21} & k_{22} & \ldots & k_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ k_{p1} & k_{p2} & \ldots & k_{pp} \end{Vmatrix}, \quad k_{\alpha\beta} = \begin{Vmatrix} k_{\alpha\beta}^{11} & k_{\alpha\beta}^{12} & \ldots & k_{\alpha\beta}^{1m_0} \\ k_{\alpha\beta}^{21} & k_{\alpha\beta}^{22} & \ldots & k_{\alpha\beta}^{2m_0} \\ \cdot & \cdot & \cdot & \cdot \\ k_{\alpha\beta}^{m_0 1} & k_{\alpha\beta}^{m_0 2} & \ldots & k_{\alpha\beta}^{m_0 m_0} \end{Vmatrix}.$$

Dans le parallélépipède à p dimensions $\overline{G} = \{0 \leqslant x_\alpha \leqslant l_\alpha, \; \alpha = 1, 2, \ldots, p\}$ à frontière Γ on étudiera le problème de Dirichlet pour le système d'équations elliptiques:

$$\begin{aligned} Lu &= \sum_{\alpha, \beta = 1}^{p} \frac{\partial}{\partial x_\alpha} \left(k_{\alpha\beta} \frac{\partial u}{\partial x_\beta} \right) = -f (x), \quad x \in G, \\ u (x) &= g (x), \quad x \in \Gamma. \end{aligned} \tag{1}$$

Si l'on passe de l'écriture vectorielle à l'écriture scalaire, le problème (1) se transcrira alors sous forme du système

$$(Lu)^s = -f^s(x), \quad x \in G,$$
$$u^s(x) = g^s(x), \quad x \in \Gamma, \quad s = 1, 2, \ldots, m_0,$$

où

$$(Lu)^s = \sum_{\alpha, \beta=1}^{p} \sum_{m=1}^{m_0} \frac{\partial}{\partial x_\alpha}\left(k_{\alpha\beta}^{sm}(x)\frac{\partial u^m}{\partial x_\beta}\right). \tag{2}$$

Admettons que la condition de forte ellipticité est remplie:

$$c_1 \sum_{\alpha=1}^{p} |\xi_\alpha|^2 \leqslant \sum_{\alpha, \beta=1}^{p} (k_{\alpha\beta}\xi_\alpha, \xi_\beta) \leqslant c_2 \sum_{\alpha=1}^{p} |\xi_\alpha|^2, \tag{3}$$

où $c_1 > 0$, $c_2 > 0$ sont des constantes indépendantes de x, $\xi_\alpha = (\xi_\alpha^1, \xi_\alpha^2, \ldots, \xi_\alpha^{m_0})$, $\alpha = 1, 2, \ldots, p$, étant des vecteurs quelconques,

$$|\xi_\alpha|^2 = \sum_{s=1}^{m_0} (\xi_\alpha^s)^2, \quad (k_{\alpha\beta}\xi_\alpha, \xi_\beta) = \sum_{s, m=1}^{m_0} k_{\alpha\beta}^{sm}\xi_\alpha^s\xi_\beta^m.$$

Notons que l'inégalité de gauche dans (3) indique que la matrice k est définie positive.

Construisons le schéma aux différences approximant le problème (1). Pour cela, dans le domaine $\bar{G}$, introduisons un maillage rectangulaire régulier

$$\bar{\omega} = \{x_i = (i_1h_1, \ldots, i_ph_p) \in \bar{G}, \ 0 \leqslant i_\alpha \leqslant N_\alpha, \quad h_\alpha N_\alpha = l_\alpha, \ \alpha = 1, 2, \ldots, p\}$$

avec frontière γ, de manière que $\bar{\omega} = \omega \cup \gamma$. On examinera sur le maillage $\bar{\omega}$ les fonctions de mailles vectorielles dont les composantes sont les fonctions de mailles p des variables discrètes, par exemple, $y = (y^1, y^2, \ldots, y^{m_0})$, avec $y^s = y^s(x_i) = y^s(i_1, i_2, \ldots, i_p)$.

Le problème discret de Dirichlet du système (1) associé au maillage $\bar{\omega}$ prend en forme vectorielle l'aspect suivant

$$\Lambda^- y = \sum_{\alpha, \beta=1}^{p} 0{,}5\,[(k_{\alpha\beta}y_{\bar{x}_\beta})_{x_\alpha} + (k_{\alpha\beta}y_{x_\beta})_{\bar{x}_\alpha}] = -\varphi(x), \quad x \in \omega,$$
$$y(x) = g(x), \quad x \in \gamma.$$

En passant à l'écriture scalaire, on obtient le système

$$(\Lambda^- y)^s = -\varphi^s(x), \quad x \in \omega, \tag{4}$$
$$y^s(x) = g^s(x), \quad x \in \gamma, \ s = 1, 2, \ldots, m_0,$$

où

$$(\Lambda^- y)^s = \sum_{\alpha, \beta=1}^{p} \sum_{m=1}^{m_0} 0{,}5\,[(k_{\alpha\beta}^{sm}y_{\bar{x}_\beta}^m)_{x_\alpha} + (k_{\alpha\beta}^{sm}y_{x_\beta}^m)_{\bar{x}_\alpha}].$$

L'opérateur Λ^-, comme au cas de l'équation elliptique scalaire, autorise une autre écriture, à savoir:

$$\Lambda^- y = \sum_{\alpha=1}^{p} 0,5\,[(k_{\alpha\alpha} y_{\bar{x}_\alpha})_{x_\alpha} + (k_{\alpha\alpha} y_{x_\alpha})_{\bar{x}_\alpha}] + \sum_{\alpha\neq\beta}^{1\div p} 0,5\,[(k_{\alpha\beta} y_{\bar{x}_\beta})_{x_\alpha} + (k_{\alpha\beta} y_{x_\beta})_{\bar{x}_\alpha}].$$

Notons que pour l'approximation de l'opérateur différentiel L il est également possible de recourir, outre Λ^-, à d'autres opérateurs de différences, par exemple

$$\Lambda^+ y = \sum_{\alpha=1}^{p} 0,5\,[(k_{\alpha\alpha} y_{\bar{x}_\alpha})_{x_\alpha} + (k_{\alpha\alpha} y_{x_\alpha})_{\bar{x}_\alpha}] + \sum_{\alpha\neq\beta}^{1\div p} 0,5\,[(k_{\alpha\beta} y_{x_\beta})_{x_\alpha} + (k_{\alpha\beta} y_{\bar{x}_\beta})_{\bar{x}_\alpha}]$$

ou bien

$$\Lambda^0 y = 0,5\,(\Lambda^- + \Lambda^+)\,y = \sum_{\alpha=1}^{p} 0,5\,[(k_{\alpha\alpha} y_{\bar{x}_\alpha})_{x_\alpha} + (k_{\alpha\alpha} y_{x_\alpha})_{\bar{x}_\alpha}] + \sum_{\alpha\neq\beta}^{1\div p} (k_{\alpha\beta} y_{\overset{\circ}{x}_\beta})_{\overset{\circ}{x}_\alpha}.$$

Introduisons l'espace H des fonctions de mailles vectorielles associées à ω et définissons-y le produit scalaire

$$(u, v) = \sum_{s=1}^{m_0} (u^s, v^s), \quad (u^s, v^s) = \sum_{x\in\omega} u^s(x)\, v^s(x)\, h_1 h_2 \ldots h_p,$$

$$u = (u^1, u^2, \ldots, u^{m_0}), \quad v = (v^1, v^2, \ldots, v^{m_0}), \quad u, v \in H.$$

Définissons l'opérateur de différences de Laplace:

$$\mathcal{R} y = \sum_{\alpha=1}^{p} y_{\bar{x}_\alpha x_\alpha}, \quad (\mathcal{R} y)^s = \sum_{\alpha=1}^{p} y^s_{\bar{x}_\alpha x_\alpha}.$$

Dans l'espace H définissons, comme d'habitude, les opérateurs A et R:

$$Ay = -\Lambda^- \overset{\circ}{y}, \quad Ry = -\mathcal{R}\overset{\circ}{y}, \quad y \in H,$$

où $\overset{\circ}{y}(x) = y(x)$ pour $x \in \omega$ et $\overset{\circ}{y}(x) = 0$ si $x \in \gamma$.

En utilisant les notations introduites et en corrigeant de façon manifeste le second membre φ de l'équation (4) aux nœuds frontières, écrivons le schéma aux différences (4) sous forme de l'équation opératorielle

$$Au = f \tag{5}$$

donnée dans l'espace hilbertien H.

En se servant de la formule de différences de Green pour les fonctions de mailles scalaires, des conditions (3) et en admettant que les conditions de symétrie sont satisfaites

$$k_{\alpha\beta}^{sm} = k_{\beta\alpha}^{ms}, \quad \alpha,\beta = 1, 2, \ldots, p, \quad s, m = 1, 2, \ldots, m_0, \quad (6)$$

on constate que les opérateurs R et A sont autoadjoints dans H et énergétiquement équivalents à constantes c_1 et c_2, c'est-à-dire qu'on est en présence d'inégalités opératorielles

$$c_1 R \leqslant A \leqslant c_2 R, \quad c_1 > 0. \quad (7)$$

Pour obtenir la solution approchée de l'équation (5), profitons de la méthode itérative implicite à deux couches avec paramètres de Tchébychev

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}} + Ay_k = f, \quad k=0, 1, \ldots, \quad y_0 \in H, \quad (8)$$

où

$$\tau_k = \frac{\tau_0}{1+\rho_0\mu_k}, \quad \mu_k \in \mathfrak{M}_n = \left\{-\cos\frac{(2i-1)\pi}{2n}, \ 1 \leqslant i \leqslant n\right\},$$
$$k = 1, 2, \ldots, n,$$
$$\tau_0 = \frac{2}{\gamma_1+\gamma_2}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2},$$
$$n \geqslant n_0(\varepsilon) = \ln(0{,}5\varepsilon)/\ln \rho_1,$$

tandis que γ_1 et γ_2 sont les constantes de l'équivalence énergétique des opérateurs autoadjoints A et B:

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \quad \gamma_1 > 0, \quad A = A^*, \quad B = B^*. \quad (9)$$

Si en guise d'opérateur B on choisit l'opérateur R défini plus haut, il s'ensuit alors de (7) que dans les inégalités (9) $\gamma_1 = c_1$ et $\gamma_2 = c_2$. Donc, le nombre d'itérations de la méthode (8) est indépendant, dans le cas considéré, du nombre de nœuds dans le maillage: $n = O(\ln(2/\varepsilon))$.

Il s'ensuit de la définition des opérateurs A et B que pour trouver y_{k+1} à partir de l'approximation précédente y_k, déjà connue, il faut résoudre le problème de différences suivant:

$$\mathcal{R}y_{k+1} = -F_k, \ x \in \omega, \quad F_k = \tau_{k+1}(\Lambda^- y_k + \varphi) - \mathcal{R}y_k,$$
$$y_{k+1} = g, \quad x \in \gamma.$$

Sous forme scalaire, ce problème s'écrit sous l'aspect du système

$$\sum_{\alpha=1}^{p} (y_{k+1}^s)_{\bar{x}_\alpha x_\alpha} = -F_k^s(x), \quad x \in \omega,$$
$$y_{k+1}^s(x) = g^s(x), \quad x \in \gamma, \quad s = 1, 2, \ldots, m_0. \quad (10)$$

Etant donné que chaque équation du système (10) peut être résolue séparément des autres équations, l'obtention de l'approximation y_{k+1} se réduit à la résolution de m_0 problèmes discrets de Dirichlet dans un parallélépipède de p dimensions associés à un maillage rectangulaire $\bar{\omega}$.

Si l'on utilise pour la résolution du problème discret de Dirichlet à p dimensions pour l'équation de Poisson la méthode de séparation des variables avec algorithme de transformation discrète rapide de Fourier, on peut montrer qu'on aura besoin de $q \approx 4pN^p \log_2 N$ $(N_1 = N_2 = \ldots = N_p = N = 2^{n_0})$ opérations arithmétiques. Par conséquent, pour résoudre le système (10) il faudra $Q_{m_0} = m_0 q$ opérations et, en tout, pour obtenir la solution du problème de différences (4) à la précision ε il est nécessaire d'effectuer $Q = nQ_{m_0} = nm_0 q = O\left(m_0 p N^p \ln \frac{2}{\varepsilon} \log_2 N\right)$ opérations arithmétiques.

Examinons maintenant la méthode itérative triangulaire alternée. Le schéma itératif prend la forme (8), où B est un opérateur factorisé $B = (E + \omega R_1)(E + \omega R_2)$, $R_1 = R_2^*$, $R_1 + R_2 = R$. Les opérateurs R_1 et R_2 se déterminent au moyen des opérateurs de différences $\mathcal{R}_1$ et $\mathcal{R}_2$ de la façon suivante: $R_\alpha y = -\mathcal{R}_\alpha \overset{\circ}{y}$, $\alpha = 1, 2$, $y(x) = \overset{\circ}{y}(x)$ pour $x \in \omega$ et $\overset{\circ}{y}(x) = 0$ pour $x \in \gamma$, où

$$\mathcal{R}_1 y = -\sum_{\alpha=1}^{p} \frac{1}{h_\alpha} y_{\bar{x}_\alpha}, \quad \mathcal{R}_2 y = \sum_{\alpha=1}^{p} \frac{1}{h_\alpha} y_{x_\alpha}.$$

Comme dans le cas scalaire, on démontre que sont satisfaites les inégalités $R \geqslant \delta E$, $R_1 R_2 \leqslant \frac{\Delta}{4} R$, où

$$\delta = \sum_{\alpha=1}^{p} \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2l_\alpha}, \quad \Delta = \sum_{\alpha=1}^{p} \frac{4}{h_\alpha^2}.$$

Il s'ensuit de la théorie générale de la méthode triangulaire alternée (voir § 1, ch. X) que, pour la valeur optimale du paramètre $\omega = \omega_0 = 2/\sqrt{\delta\Delta}$, on obtient les inégalités opératorielles

$$\overset{\circ}{\gamma}_1 B \leqslant R \leqslant \overset{\circ}{\gamma}_2 B, \quad \overset{\circ}{\gamma}_1 > 0, \tag{11}$$

où $\overset{\circ}{\gamma}_1 = \dfrac{\delta}{2(1+\sqrt{\eta})}$, $\overset{\circ}{\gamma}_2 = \dfrac{\delta}{4\sqrt{\eta}}$, $\eta = \dfrac{\delta}{\Delta}$.

En comparant (7), (9) et (11), on trouve que les opérateurs A et B vérifient les inégalités (9) avec $\gamma_1 = c_1 \overset{\circ}{\gamma}_1$ et $\gamma_2 = c_2 \overset{\circ}{\gamma}_2$.

En se servant pour le schéma (8) du jeu de paramètres τ_k de Tchébychev on constate que la méthode itérative triangulaire

alternée construite exige $n = O \frac{1}{\sqrt{|h|}}\left(\sqrt{\frac{c_2}{c_1}} \ln \varepsilon \frac{2}{\varepsilon}\right)$ itérations, où $|h|^2 = h_1^2 + h_2^2 + \ldots + h_p^2$. Vu que le passage de y_k à y_{k+1} s'effectue suivant des formules explicites en $O(m_0 N_1 N_2 \ldots N_p)$ opérations arithmétiques, le nombre total d'opérations qu'il est nécessaire de dépenser pour obtenir la solution du problème (4) à la précision ε est estimé à

$$Q = O\left(m_0 N^{p+0,5} \sqrt{\frac{c_2}{c_1}} \ln \frac{2}{\varepsilon}\right),$$

si $l_1 = l_2 = \ldots = l_p$, $N_1 = N_2 = \ldots = N_p = N$.

Notons en conclusion que les méthodes itératives passées en revue plus haut convergent dans l'espace énergétique H_D, où en guise d'opérateur D il est possible de choisir l'un des opérateurs A, B ou $AB^{-1}A$.

2. Système d'équations de la théorie de l'élasticité. Prenons le système d'équations de la théorie de l'élasticité stationnaire (équations de Lamé)

$$Lu = \mu \Delta u + (\lambda + \mu) \operatorname{grad} \operatorname{div} u = -f(x), \tag{12}$$

où $u = (u^1, u^2, \ldots, u^p)$, $f = (f^1, f^2, \ldots, f^p)$, $x = (x_1, x_2, \ldots \ldots, x_p)$, $\lambda > 0$ et $\mu > 0$ étant les paramètres de Lamé.

Ecrivons l'équation (12) sous forme du système

$$(Lu)^s = \mu \sum_{\alpha=1}^{p} \frac{\partial^2 u^s}{\partial x_\alpha^2} + (\lambda + \mu) \sum_{\beta=1}^{p} \frac{\partial^2 u^\beta}{\partial x_\beta \partial x_s} = -f^s, \quad s = 1, 2, \ldots, p. \tag{13}$$

Pour $p = 2$ le système (13) peut être écrit sous la forme

$$(\lambda + 2\mu) \frac{\partial^2 u^1}{\partial x_1^2} + \mu \frac{\partial^2 u^1}{\partial x_2^2} + (\lambda + \mu) \frac{\partial^2 u^2}{\partial x_1 \partial x_2} = -f^1(x_1, x_2),$$

$$(\lambda + \mu) \frac{\partial^2 u^1}{\partial x_1 \partial x_2} + \mu \frac{\partial^2 u^2}{\partial x_1^2} + (\lambda + 2\mu) \frac{\partial^2 u^2}{\partial x_2^2} = -f^2(x_1, x_2).$$

Ce système décrit l'équilibre d'un solide élastique homogène et isotrope pour le cas d'une déformation plane. Les fonctions inconnues $u^1(x_1, x_2)$ et $u^2(x_1, x_2)$ ont la signification de déplacements du point dans les directions des axes Ox_1 et Ox_2 respectivement.

Pour le système (12) on peut poser le problème de la recherche du vecteur $u(x)$ qui satisfait à l'équation (12) dans le domaine G et adoptant à la frontière Γ les valeurs données

$$u(x) = g(x), \quad x \in \Gamma. \tag{14}$$

En confrontant (13) avec (2) on trouve que le système (12), (14) peut être écrit sous la forme (1), où $m_0 = p$,

$$k_{\alpha\beta}^{sm} = \mu \delta_{\alpha\beta} \delta_{sm} + (\lambda + \mu) [\theta \delta_{\alpha s} \delta_{\beta m} + (1 - \theta) \delta_{\alpha m} \delta_{\beta s}], \tag{15}$$

tandis que θ est une constante arbitraire, $\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$
En effet, en portant (15) dans (2), on a

$$(Lu)^s = \sum_{\alpha,\,\beta=1}^{p} \sum_{m=1}^{p} \frac{\partial}{\partial x_\alpha}\left(k_{\alpha\beta}^{sm} \frac{\partial u^m}{\partial x_\beta}\right) = \mu \sum_{\alpha,\,\beta=1}^{p} \sum_{m=1}^{p} \delta_{\alpha\beta}\delta_{sm} \frac{\partial^2 u^m}{\partial x_\alpha\,\partial x_\beta} +$$

$$+ (\lambda+\mu)\left[\theta \sum_{\alpha,\,\beta=1}^{p} \sum_{m=1}^{p} \delta_{\alpha s}\delta_{\beta m} \frac{\partial^2 u^m}{\partial x_\alpha\,\partial x_\beta} +\right.$$

$$\left. + (1-\theta) \sum_{\alpha,\,\beta=1}^{p} \sum_{m=1}^{p} \delta_{\alpha m}\delta_{\beta s} \frac{\partial^2 u^m}{\partial x_\alpha\,\partial x_\beta}\right] =$$

$$= \mu \sum_{\alpha=1}^{p} \frac{\partial^2 u^s}{\partial x_\alpha^2} + (\lambda+\mu)\left[\theta \sum_{\beta=1}^{p} \frac{\partial^2 u^\beta}{\partial x_s\,\partial x_\beta} + (1-\theta) \sum_{\alpha=1}^{p} \frac{\partial^2 u^\alpha}{\partial x_\alpha\,\partial x_s}\right] =$$

$$= \mu \sum_{\alpha+1}^{p} \frac{\partial^2 u^s}{\partial x_\alpha^2} + (\lambda+\mu) \sum_{\beta=1}^{p} \frac{\partial^2 u^\beta}{\partial x_s\,\partial x_\beta}.$$

La proposition est démontrée.

Cherchons maintenant les constantes c_1 et c_2 des inégalités (3). Montrons que $c_1 = \mu$. On a

$$\sum_{s,\,m=1}^{p} \sum_{\alpha,\,\beta=1}^{p} k_{\alpha\beta}^{sm}\xi_\alpha^s\xi_\beta^m = \mu \sum_{\alpha,\,s=1}^{p} (\xi_\alpha^s)^2 +$$

$$+ (\lambda+\mu)\left[\theta \sum_{\alpha,\,s=1}^{p} \xi_\alpha^\alpha\xi_s^s + (1-\theta) \sum_{\alpha,\,s=1}^{p} \xi_\alpha^s\xi_s^\alpha\right] =$$

$$= \mu \sum_{\alpha,\,s=1}^{p} (\xi_\alpha^s)^2 + (\lambda+\mu)\left[\theta\left(\sum_{\alpha=1}^{p} \xi_\alpha^\alpha\right)^2 + (1-\theta) \sum_{\alpha,\,s=1}^{p} \xi_\alpha^s\xi_s^\alpha\right]. \qquad (16)$$

En posant ici $\theta = 1$, il vient

$$\sum_{\alpha,\,\beta=1}^{p} (k_{\alpha\beta}\xi_\alpha.\,\xi_\beta) = \mu \sum_{\alpha=1}^{p} |\,\xi_\alpha\,|^2 + (\lambda+\mu)\left(\sum_{\alpha=1}^{p} \xi_\alpha^\alpha\right)^2 \geqslant \mu \sum_{\alpha=1}^{p} |\,\xi_\alpha\,|^2.$$

On montre sans peine également que $c_2 = \lambda + 2\mu$. En posant dans (16) $\theta = 0$ et en utilisant l'inégalité de Cauchy-Bouniakovski, on obtient

$$\sum_{\alpha,\,\beta=1}^{p} (k_{\alpha\beta}\xi_\alpha,\, \xi_\beta) = \mu \sum_{\alpha=1}^{p} |\,\xi_\alpha\,|^2 + (\lambda+\mu) \sum_{\alpha,\,s=1}^{p} \xi_\alpha^s\xi_s^\alpha \leqslant$$

$$\leqslant \mu \sum_{\alpha=1}^{p} |\xi_\alpha|^2 + \frac{(\lambda+\mu)}{2}\left[\sum_{\alpha,\, s=1}^{p} (\xi_\alpha^s)^2 + \sum_{\alpha,\, s=1}^{p} (\xi_s^\alpha)^2\right] =$$

$$= \mu \sum_{\alpha=1}^{p} |\xi_\alpha|^2 + (\lambda+\mu) \sum_{\alpha,\, s=1}^{p} (\xi_\alpha^s)^2 = (\lambda+2\mu) \sum_{\alpha=1}^{p} |\xi_\alpha|^2.$$

Construisons maintenant le schéma aux différences approximant le problème (12), (14). En portant (15) dans le schéma aux différences (4), il vient

$$(\Lambda^- y)^s = \mu \sum_{\alpha=1}^{p} y^s_{\bar{x}_\alpha x_\alpha} + 0{,}5\,(\lambda+\mu) \sum_{\beta=1}^{p} (y^\beta_{\bar{x}_\beta x_s} + y^\beta_{x_\beta \bar{x}_s}) = -\varphi^s, \quad x \in \omega, \tag{17}$$
$$y^s(x) = g^s(x), \quad x \in \gamma, \quad s = 1, 2, \ldots, p,$$

où $\bar{\omega} = \omega \cup \gamma$ est le maillage introduit au point 1.

Il reste à déterminer les opérateurs A et R, comme on l'a fait au point 1. La condition de symétrie (6) est satisfaite, aussi, en utilisant la première formule de différences de Green, obtient-on les opérateurs A et R autoadjoints dans H, de plus on a les inégalités $c_1 R \leqslant A \leqslant c_2 R$, où $c_1 = \mu$, $c_2 = \lambda + 2\mu$.

Les raisonnements subséquents coïncident ici avec ceux menés au point 1. C'est ainsi que la méthode itérative (8), avec $B = R$ et des paramètres de Tchébychev τ_k, présente l'estimation suivante pour le nombre d'itérations:

$$n \geqslant n_0(\varepsilon) = \frac{\ln 0{,}5\varepsilon}{\ln \rho_1}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{c_1}{c_2} = \frac{\mu}{\lambda+2\mu},$$

tandis que la méthode triangulaire alternée, construite sur la base du régularisateur R, se caractérise par la même estimation, où

$$\xi = \frac{\mu}{\lambda+2\mu}\,\frac{1\sqrt{\eta}}{1+\sqrt{\eta}}, \quad \eta = \frac{\delta}{\Delta},$$

$$\delta = \sum_{\alpha=1}^{p} \frac{4}{h_\alpha^2} \sin^2 \frac{\pi h_\alpha}{2 l_\alpha}, \quad \Delta = \sum_{\alpha=1}^{p} \frac{4}{h_\alpha^2}.$$

Ainsi donc pour la méthode triangulaire alternée le nombre d'itérations est proportionnel à $\sqrt{\frac{\lambda+2\mu}{\mu}} = \sqrt{2+\frac{\lambda}{\mu}}$:

$$n_0(\varepsilon) = \sqrt{2+\frac{\lambda}{\mu}}\, n_0^*(\varepsilon),$$

où $n_0^*(\varepsilon)$ est le nombre d'itérations nécessaire à la résolution de l'équation aux différences de Poisson à p dimensions par la méthode triangulaire alternée.

CHAPITRE XV

MÉTHODES DE RÉSOLUTION DES ÉQUATIONS ELLIPTIQUES EN COORDONNÉES CURVILIGNES ORTHOGONALES

Dans ce chapitre on étudie des exemples de résolution des problèmes de différences approximant les problèmes aux limites pour des équations elliptiques dans des systèmes de coordonnées curvilignes. On établit les conditions d'applicabilité des méthodes directes et itératives, en particulier de la méthode des directions alternées, aux problèmes en coordonnées cylindriques et polaires.

Dans le § 1 on montre comment se posent les problèmes aux limites pour des équations différentielles. Le § 2 est consacré à l'exposé des méthodes directes et itératives de résolution des problèmes de différences en géométrie (r, z), de même que des problèmes sur la surface du cylindre. Dans le § 3 sont étudiées les méthodes de résolution des problèmes de différences dans le cercle, l'anneau et le secteur annulaire.

§ 1. Position des problèmes aux limites pour des équations différentielles

1. Equations elliptiques dans le système de coordonnées cylindriques. Soit donnée l'équation de Poisson

$$Lu = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \frac{\partial^2 u}{\partial x_3^2} = -f(x), \quad x = (x_1, x_2, x_3). \tag{1}$$

S'il s'agit avec cette équation d'obtenir la solution dans un cylindre circulaire fini ou dans un tube circulaire, il est naturel de l'étudier en coordonnées cylindriques. Dans ce système de coordonnées l'équation de Poisson (1) prend la forme

$$L_{r\varphi z} u = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 u}{\partial \varphi^2} + \frac{\partial^2 u}{\partial z^2} + - f(r, \varphi, z), \tag{2}$$

où $r = \sqrt{x_1^2 + x_2^2}$, $\operatorname{tg} \varphi = x_2/x_1$, $z = x_3$.

L'équation (1) décrit, par exemple, une distribution stationnaire de la température $u = u(x_1, x_2, x_3)$ dans un milieu homogène. Si le milieu n'est pas homogène, mais isotrope, au lieu de (1) il faut étudier l'équation

$$Lu = \operatorname{div}(k \operatorname{grad} u) = \sum_{\alpha=1}^{3} \frac{\partial}{\partial x_\alpha}\left(k(x)\frac{\partial}{\partial x_\alpha}\right) = -f(x), \tag{3}$$

à laquelle dans le système (r, φ, z) correspond l'équation

$$L_{r\varphi z}u = \frac{1}{r}\frac{\partial}{\partial r}\left(rk\frac{\partial u}{\partial r}\right) + \frac{1}{r^2}\frac{\partial}{\partial \varphi}\left(k\frac{\partial u}{\partial \varphi}\right) + \frac{\partial}{\partial z}\left(k\frac{\partial u}{\partial z}\right) = -f. \quad (4)$$

Si le milieu est anisotrope, c'est-à-dire que le coefficient de conductibilité thermique est fonction non seulement du point, mais également de la direction, on aura alors au lieu de (3) une équation aux dérivées mixtes

$$Lu = \sum_{\alpha,\ \beta=1}^{3} \frac{\partial}{\partial x_\alpha}\left(k_{\alpha\beta}\frac{\partial u}{\partial x_\beta}\right) = -f(x). \quad (5)$$

L'équation (5) en coordonnées cylindriques correspond à l'équation

$$L_{r\varphi z}u = \frac{1}{r}\frac{\partial}{\partial r}\left[r\left(\bar{k}_{11}\frac{\partial u}{\partial r} + \frac{\bar{k}_{12}}{r}\frac{\partial u}{\partial \varphi} + \bar{k}_{13}\frac{\partial u}{\partial z}\right)\right] +$$
$$+ \frac{1}{r}\frac{\partial}{\partial \varphi}\left(\bar{k}_{21}\frac{\partial u}{\partial r} + \frac{\bar{k}_{22}}{r}\frac{\partial u}{\partial \varphi} + \bar{k}_{23}\frac{\partial u}{\partial z}\right) +$$
$$+ \frac{\partial}{\partial z}\left(\bar{k}_{31}\frac{\partial u}{\partial r} + \frac{\bar{k}_{32}}{r}\frac{\partial u}{\partial \varphi} + \bar{k}_{33}\frac{\partial u}{\partial z}\right) = -f(r, \varphi, z), \quad (6)$$

où les coefficients $\bar{k}_{\alpha\beta}$ s'expriment au moyen de $k_{\alpha\beta}$ suivant les formules:

$$\bar{k}_{11} = k_{11}\cos^2\varphi + (k_{12} + k_{21})\sin\varphi\cos\varphi + k_{22}\sin^2\varphi,$$
$$\bar{k}_{12} = k_{12}\cos^2\varphi + (k_{22} - k_{11})\sin\varphi\cos\varphi - k_{21}\sin^2\varphi,$$
$$\bar{k}_{21} = k_{21}\cos^2\varphi + (k_{22} - k_{11})\sin\varphi\cos\varphi - k_{12}\sin^2\varphi,$$
$$\bar{k}_{22} = k_{11}\sin^2\varphi - (k_{12} + k_{21})\sin\varphi\cos\varphi + k_{22}\cos^2\varphi,$$
$$\bar{k}_{13} = k_{13}\cos\varphi + k_{23}\sin\varphi,\quad \bar{k}_{23} = k_{23}\cos\varphi - k_{13}\sin\varphi,$$
$$\bar{k}_{31} = k_{31}\cos\varphi + k_{32}\sin\varphi,\quad \bar{k}_{32} = k_{32}\cos\varphi - k_{31}\sin\varphi,$$
$$\bar{k}_{33} = k_{33}.$$

L'équation (6) est appelée *équation à dérivées mixtes en système de coordonnées cylindriques*. Si $\bar{k}_{\alpha\beta} = 0$ pour $\alpha \neq \beta$, (6) prend alors la forme

$$L_{r\varphi z}u = \frac{1}{r}\frac{\partial}{\partial r}\left(r\bar{k}_1\frac{\partial u}{\partial r}\right) + \frac{1}{r^2}\frac{\partial}{\partial \varphi}\left(\bar{k}_2\frac{\partial u}{\partial \varphi}\right) + \frac{\partial}{\partial z}\left(\bar{k}_3\frac{\partial u}{\partial z}\right) = -f, \quad (7)$$

où $\bar{k}_\alpha = \bar{k}_{\alpha\alpha}$, $\alpha = 1, 2, 3$ et porte le nom d'équation sans dérivées mixtes.

Notons que si $k_{\alpha\beta} = k_{\beta\alpha}$, on a aussi $\bar{k}_{\alpha\beta} = \bar{k}_{\beta\alpha}$ et inversement. Les équations susmentionnées (2) et (4) sont des cas particuliers de l'équation (7) correspondant à $\bar{k}_\alpha \equiv 1$ et à $\bar{k}_\alpha \equiv k$.

L'équation (5) devient fortement elliptique au cas où il existe une constante $c_1 > 0$ qui, pour tous ξ_1, ξ_2 et ξ_3, vérifie l'inégalité

$$\sum_{\alpha, \beta=1}^{3} k_{\alpha\beta}(x)\, \xi_\alpha \xi_\beta \geqslant c_1 \sum_{\alpha=1}^{3} \xi_\alpha^2. \tag{8}$$

Si on effectue dans (8) la substitution en posant

$$\xi_1 = \bar{\xi}_1 \cos \varphi - \bar{\xi}^2 \sin \varphi, \quad \xi_2 = \bar{\xi}_1 \sin \varphi + \bar{\xi}_2 \cos \varphi, \quad \xi_3 = \bar{\xi}_3,$$

l'inégalité (8) devient alors de la forme

$$\sum_{\alpha, \beta=1}^{3} \bar{k}_{\alpha\beta} \bar{\xi}_\alpha \bar{\xi}_\beta \geqslant c_1 \sum_{\alpha=1}^{3} \bar{\xi}_\alpha^2. \tag{9}$$

En pratique on rencontre le plus souvent deux cas.

A) Dans le cas de symétrie axiale les coefficients et le second membre de l'équation, comme la solution elle-même, ne dépendent pas de l'angle φ. De plus, l'équation (6) se simplifie

$$L_{rz} u = \frac{1}{r} \frac{\partial}{\partial r} \left[r \left(\bar{k}_{11} \frac{\partial u}{\partial r} + \bar{k}_{13} \frac{\partial u}{\partial z} \right) \right] + \\ + \frac{\partial}{\partial z} \left(\bar{k}_{31} \frac{\partial u}{\partial r} + \bar{k}_{33} \frac{\partial u}{\partial z} \right) = -f(r, z), \tag{10}$$

tandis qu'en l'absence de dérivées mixtes l'équation correspondant à (7) prend la forme

$$L_{rz} u = \frac{1}{r} \frac{\partial}{\partial r} \left(r \bar{k}_1 \frac{\partial u}{\partial r} \right) + \frac{\partial}{\partial z} \left(\bar{k}_3 \frac{\partial u}{\partial z} \right) = -f(r, z). \tag{11}$$

B) Dans le cas plan les coefficients, le second membre et la solution de l'équation (6) sont indépendants de z, et, par suite, l'équation (6) prend la forme

$$L_{r\varphi} u = \frac{1}{r} \frac{\partial}{\partial r} \left[r \left(\bar{k}_{11} \frac{\partial u}{\partial r} + \frac{\bar{k}_{12}}{r} \frac{\partial u}{\partial \varphi} \right) \right] + \\ + \frac{1}{r} \frac{\partial}{\partial \varphi} \left(\bar{k}_{21} \frac{\partial u}{\partial r} + \frac{\bar{k}_{22}}{r} \frac{\partial u}{\partial \varphi} \right) = -f(r, \varphi). \tag{12}$$

Si les dérivées mixtes manquent, l'équation acquiert la forme

$$L_{r\varphi} u = \frac{1}{r} \frac{\partial}{\partial r} \left(r \bar{k}_1 \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial}{\partial \varphi} \left(\bar{k}_2 \frac{\partial u}{\partial \varphi} \right) = -f(r, \varphi). \tag{13}$$

Pour le cas plan on dit que (12) et (13) sont des équations elliptiques en coordonnées polaires.

Remarquons que pour $\bar{k}_\alpha \equiv 1$, $\alpha = 1, 2, 3$ les formules (11) et (13) décrivent l'équation de Poisson en coordonnées (r, z) et (r, φ).

On est parfois obligé de résoudre l'équation de Poisson ou une équation elliptique plus générale sur la surface d'un cylindre de

rayon R. Dans ce cas

$$L_{\varphi z} = \frac{1}{R}\frac{\partial}{\partial \varphi}\left(\frac{\bar{k}_{22}}{R}\frac{\partial u}{\partial \varphi} + + \bar{k}_{23}\frac{\partial u}{\partial z}\right) + \\ + \frac{\partial}{\partial z}\left(\frac{\bar{k}_{32}}{R}\frac{\partial u}{\partial \varphi} + \bar{k}_{33}\frac{\partial u}{\partial z}\right) = -f(\varphi, z), \quad (14)$$

tandis que l'équation (7) sans dérivées mixtes prend la forme

$$L_{\varphi z}u = \frac{1}{R^2}\frac{\partial}{\partial \varphi}\left(\bar{k}_2\frac{\partial u}{\partial \varphi}\right) + \frac{\partial}{\partial z}\left(\bar{k}_3\frac{\partial u}{\partial z}\right) = -f(\varphi, z). \quad (14')$$

Notons que la substitution $\varphi' = R\varphi$ permet de réduire ces équations à des équations elliptiques ordinaires à coefficients variables.

2. Problèmes aux limites pour équations dans un système de coordonnées cylindriques. Examinons d'abord le cas de *symétrie axiale*. La solution ne dépendant pas de l'angle φ, le domaine où est recherchée la solution en coordonnées cylindriques (r, z) constitue un rectangle $\bar{G} = \{l_1 \leqslant r \leqslant L_1,\ l_3 \leqslant z \leqslant L_3,\ l_1 \geqslant 0\}$. Si le domaine initial est un cylindre annulaire (creux), alors $l_1 > 0$.

Posons les problèmes aux limites pour l'équation (10) dans le rectangle $\bar{G}$. Dans le domaine G on donne l'équation (10) et sur les côtés $r = L_1$, $z = l_3$ et $z = L_3$ l'une des conditions aux limites de première, de deuxième ou de troisième espèce. Ainsi, les conditions aux limites de troisième espèce sont de la forme

$$\begin{aligned} -\bar{k}_{11}\frac{\partial u}{\partial r} - \bar{k}_{13}\frac{\partial u}{\partial z} &= \varkappa_1^+ u - g_1^+(z), \quad r = L_1, \\ \bar{k}_{31}\frac{\partial u}{\partial r} + \bar{k}_{33}\frac{\partial u}{\partial z} &= \varkappa_3^- u - g_3^-(r), \quad z = l_3, \\ -\bar{k}_{31}\frac{\partial u}{\partial r} - \bar{k}_{33}\frac{\partial u}{\partial z} &= \varkappa_3^+ u - g_3^+(r), \quad z = L_3. \end{aligned} \quad (15)$$

Pour $l_1 = 0$ l'équation (10) présente une singularité sur l'axe $r = 0$. On s'intéresse dans ce cas à une solution limitée. Si $l_1 > 0$, au côté $r = l_1$ peut être imposée une des conditions aux limites de première, de deuxième ou de troisième espèce. C'est ainsi que la condition de troisième espèce est de la forme

$$\bar{k}_{11}\frac{\partial u}{\partial r} + \bar{k}_{13}\frac{\partial u}{\partial z} = \varkappa_1^- u - g_1^-(z), \quad r = l_1 > 0. \quad (16)$$

Si $l_1 = 0$, la solution limitée se distingue par la condition

$$\lim_{r \to 0} r\left(\bar{k}_{11}\frac{\partial u}{\partial r} + \bar{k}_{13}\frac{\partial u}{\partial z}\right) = 0. \quad (17)$$

Dans les conditions (15), (16) $\varkappa_1^{\pm}(z)$ et $\varkappa_3^{\pm}(r)$ sont des fonctions non négatives. Si à la frontière du rectangle $\overline{G}$ sont imposées les conditions aux limites de deuxième espèce $(\varkappa_1^{\pm} \equiv 0)$, le problème (10) (15), (16) n'admet une solution qu'à la satisfaction de condition

$$\int_{l_1}^{L_1}\int_{l_3}^{L_3} rf(r, z)\,dr\,dz + \int_{l_3}^{L_3} [L_1 g_1^+(z) + l_1 g_3^-(z)]\,dz + \\ + \int_{l_1}^{L_1} r\,[g_3^+(z) + g_3^-(r)]\,dr = 0. \qquad (18)$$

Dans ce cas la solution n'est pas unique et se définit à la précision de la constante près, c'est-à-dire on a $u(r, z) = u_0(r, z) + \text{const}$, où $u_0(r, z)$ est une solution quelconque.

Examinons maintenant *l'équation* (14) *sur la surface d'un cylindre*. En coordonnées (φ, z) le domaine où est recherchée la solution est le rectangle $\overline{G} = \{l_2 \leqslant \varphi \leqslant L_2,\ l_3 \leqslant z \leqslant L_3,\ L_2 - l_2 \leqslant 2\pi\}$.

Aux côtés $z = l_3$ et $z = L_3$ peuvent être imposées les conditions aux limites de première, de deuxième ou de troisième espèce, par exemple

$$\begin{aligned} \frac{1}{R}\overline{k}_{32}\frac{\partial u}{\partial \varphi} + \overline{k}_{33}\frac{\partial u}{\partial z} &= \varkappa_3^- u - g_3^-(\varphi), \quad z = l_3, \\ -\frac{1}{R}\overline{k}_{32}\frac{\partial u}{\partial \varphi} - \overline{k}_{33}\frac{\partial u}{\partial z} &= \varkappa_3^+ u - g_3^+(\varphi), \quad z = L_3. \end{aligned} \qquad (19)$$

Les conditions aux limites de ce type peuvent être imposées aux côtés $\varphi = l_2$ et $\varphi = L_2$ au cas où la surface n'est pas fermée ($L_2 - l_2 < 2\pi$). C'est ainsi que les conditions aux limites de troisième espèce sont de la forme

$$\begin{aligned} \frac{1}{R^2}\overline{k}_{22}\frac{\partial u}{\partial \varphi} + \frac{1}{R}\overline{k}_{23}\frac{\partial u}{\partial z} &= \varkappa_2^- u - g_2^-(z), \quad \varphi = l_2, \\ -\frac{1}{R^2}\overline{k}_{22}\frac{\partial u}{\partial \varphi} - \frac{1}{R}\overline{k}_{23}\frac{\partial u}{\partial z} &= \varkappa_2^+ u - g_2^+(z), \quad \varphi = L_2. \end{aligned} \qquad (20)$$

Dans ce cas $\varkappa_2^{\pm}(z) \geqslant 0$ et $\varkappa_3^{\pm}(\varphi) \geqslant 0$.

La condition de résolubilité du problème (14), (19), (20) avec $\varkappa_\alpha^{\pm} \equiv 0$ acquiert la forme

$$\int_{l_2}^{L_2}\int_{l_3}^{L_3} f(\varphi, z)\,d\varphi\,dz + \int_{l_3}^{L_3} [g_2^+(z) + g_2^-(z)]\,dz + \int_{l_2}^{L_2} [g_3^+(\varphi) + g_3^-(\varphi)]\,d\varphi = 0.$$

Si la surface est fermée ($L_2 - l_2 = 2\pi$), les côtés $\varphi = l_2$ et $\varphi = L_2$ sont identifiés et l'on pose le problème de la recherche de la solution périodique (de période 2π) de l'équation (14) satisfaisant sur les côtés $z = l_3$ et $z = L_3$ à l'une des conditions susmentionnées. Si, de plus, dans les conditions (19) $\varkappa_3^{\pm} \equiv 0$, la condition de résolubilité

(à la précision de la constante près) du problème considéré prend la forme

$$\int_{l_2}^{L_2}\int_{l_3}^{L_3} f(\varphi, z)\,d\varphi\,dz + \int_{l_2}^{L_2} [g_3^+(\varphi) + g_3^-(\varphi)]\,d\varphi = 0.$$

Formulons à présent *les positions des problèmes aux limites pour l'équation* (12) *donnée en coordonnées polaires* pour le cas où le domaine considéré en coordonnées cartésiennes variables est un cercle, un anneau ou un secteur annulaire. En coordonnées (r, φ), aux domaines mentionnés correspond le rectangle $\overline{G} = \{l_1 \leqslant r \leqslant L_1,\ l_2 \leqslant \leqslant \varphi \leqslant L_2,\ l_1 \geqslant 0,\ L_2 - l_2 \leqslant 2\pi\}$.

Supposons d'abord que *le domaine initial est un cercle*. L'équation (12) est donnée dans G; pour $r = L_1$ on impose l'une des conditions aux limites de première, de deuxième ou de troisième espèce. Par exemple, la condition aux limites de troisième espèce est de la forme

$$-\overline{k}_{11}\frac{\partial u}{\partial r} - \frac{\overline{k}_{12}}{r}\frac{\partial u}{\partial \varphi} = \varkappa_1^+ u - g_1^+(\varphi), \quad r = L_1. \tag{21}$$

Pour que le problème (12), (21) soit correct il faut imposer une condition supplémentaire au centre du cercle. On recherche habituellement la solution limitée pour $r = 0$. Cette solution satisfait à la condition

$$\lim_{r\to 0} r\left(\overline{k}_{11}\frac{\partial u}{\partial r} + \frac{\overline{k}_{12}}{r}\frac{\partial u}{\partial \varphi}\right) = 0. \tag{22}$$

Vu qu'en coordonnées polaires le point $r = 0$ du plan (x_1, x_2) a une coordonnée arbitraire φ, tous les points du côté du rectangle $\overline{G}$ sont, pour $r = 0$, identiques. De plus, $u(0, \varphi) = u_0 = \text{const}$ pour $l_2 \leqslant \leqslant \varphi \leqslant L_2$ en vertu de la continuité de la solution.

Ensuite, les côtés $\varphi = l_2$ et $\varphi = L_2$ sont rendus identiques et l'on pose le problème de la recherche de la solution périodique de période 2π de l'équation (12) qui satisfait aux conditions susmentionnées.

Dans le cas où, pour $r = L_1$, est posée la condition aux limites (21) de deuxième espèce avec $\varkappa_1^+(\varphi) \equiv 0$, la solution du problème existe si est remplie la condition

$$\int_0^{2\pi}\int_0^{L_1} rf(r, \varphi)\,dr\,d\varphi + L_1\int_0^{2\pi} g_1^+(\varphi)\,d\varphi = 0. \tag{23}$$

La solution dans ce cas n'est pas unique et est définie à la précision de la constante près.

Posons maintenant que *le domaine de départ est un anneau*, c'est-à-dire $l_1 > 0$. On cherche alors la solution périodique de période 2π de l'équation (12) satisfaisant sur les côtés $r = l_1$ et $r = L_1$ à l'une

des conditions aux limites de première, de deuxième ou de troisième espèce. Donnons l'aspect de la condition aux limites de troisième espèce sur la face interne de l'anneau

$$\bar{k}_{11}\frac{\partial u}{\partial r}+\frac{\bar{k}_{12}}{r}\frac{\partial u}{\partial \varphi}=\varkappa_1^- u-g_1^-(\varphi), \quad r=l_1, \tag{24}$$

où $\varkappa_1^-(\varphi)\geqslant 0$.

Si sont données les conditions aux limites de deuxième espèce (21), (24) avec $\varkappa_1^{\pm}(\varphi)\equiv 0$, la solution du problème posé existe si est remplie la condition

$$\int_0^{2\pi}\int_{l_1}^{L_1} rf(r,\varphi)\,dr\,d\varphi+\int_0^{2\pi}[L_1g_1^+(\varphi)+l_1g_1^-(\varphi)]\,d\varphi=0. \tag{25}$$

Dans ce cas la solution est définie à la précision de la constante près.

Si *le domaine est un secteur annulaire* ($l_1>0$, $L_2-l_2<2\pi$), on pose le problème de la recherche de la solution de l'équation (12) satisfaisant sur les côtés du rectangle $\bar{G}$ à l'une des conditions de première, de deuxième ou de troisième espèce, en particulier aux conditions (21) (24) pour $r=L_1$ et $r=l_1$ et aux conditions aux limites de troisième espèce

$$\begin{aligned}\bar{k}_{21}\frac{\partial u}{\partial r}+\frac{\bar{k}_{22}}{\partial r}\frac{\partial u}{\partial \varphi}&=\varkappa_2^- u-g_2^-(r), \quad \varphi=l_2,\\ -\bar{k}_{21}\frac{\partial u}{\partial r}-\frac{\bar{k}_{22}}{\partial r}\frac{\partial u}{\partial \varphi}&=\varkappa_2^+ u-g_2^+(r), \quad \varphi=L_2,\end{aligned} \tag{26}$$

pour $\varphi=l_2$ et $\varphi=L_2$, $\varkappa_2^{\pm}(r)\geqslant 0$.

Si sont données les conditions aux limites de deuxième espèce (21), (24), (26) avec $\varkappa_1^{\pm}(\varphi)\equiv 0$, $\varkappa_2^{\pm}(r)\equiv 0$, la solution du problème existe au cas où est remplie la condition

$$\int_{l_2}^{L_2}\int_{l_1}^{L_1} rf(r,\varphi)\,dr\,d\varphi+\int_{l_2}^{L_2}(L_1g_1^++l_1g_1^-)\,d\varphi+\int_{l_1}^{L_1}(g_2^-+g_2^+)\,dr=0. \tag{27}$$

La solution dans ce cas n'est pas unique et se définit à la précision de la constante près.

§ 2. Résolution des problèmes de différences en coordonnées cylindriques

1. Schémas aux différences sans dérivées mixtes au cas d'une symétrie axiale. Examinons les problèmes aux limites pour des équations elliptiques sans dérivées mixtes en coordonnées cylindriques au cas d'une symétrie axiale.

Il s'agit de trouver dans le rectangle $\bar{G}=\{l_1\leqslant r\leqslant L_1,\ l_3\leqslant z\leqslant L_3,\ l_1\geqslant 0\}$ la solution de l'équation

$$\frac{1}{r}\frac{\partial}{\partial r}\left(rk_1\frac{\partial u}{\partial r}\right)+\frac{\partial u}{\partial z}\left(k_3\frac{\partial u}{\partial z}\right)-qu=-f(r, z), \quad (r, z)\in G, \tag{1}$$

satisfaisant à la frontière du rectangle $\overline{G}$ aux conditions aux limites suivantes:

1) sur le côté $r = l_1$, $l_3 \leqslant z \leqslant L_3$,

$$u(r, z) = g_1^-(z), \text{ si } l_1 > 0 \tag{2}$$

ou

$$\begin{aligned} &k_1\frac{\partial u}{\partial r}=\varkappa_1^- u-g_1^-(z), \quad \text{si} \quad l_1>0,\\ &\lim_{r\to 0} rk_1\frac{\partial u}{\partial r}=0, \qquad \text{si} \quad l_1=0; \end{aligned} \tag{3}$$

2) sur le côté $r=L_1$, $l_3\leqslant z\leqslant L_3$,

$$u(r, z)=g_1^+(z) \tag{4}$$

ou

$$-k_1\frac{\partial u}{\partial r}=\varkappa_1^+u-g_1^+(z); \tag{5}$$

3) sur le côté $z=l_3$, $l_1\leqslant r\leqslant L_1$,

$$u(r, z)=g_3^-(r) \tag{6}$$

ou

$$k_3\frac{\partial u}{\partial z}=\varkappa_3^-u-g_3^-(r); \tag{7}$$

4) sur le côté $z=L_3$, $l_1\leqslant r\leqslant L_1$,

$$u(r, z)=g_3^+(r) \tag{8}$$

ou

$$-k_3\frac{\partial u}{\partial z}=\varkappa_3^+u-g_3^+(r). \tag{9}$$

On admet que les coefficients satisfont aux conditions

$$k_1(r, z)\geqslant c_1>0, \quad k_3(r, z)\geqslant c_1>0, \quad q(r, z)\geqslant 0,$$

$$\varkappa_1^\pm(z)\geqslant 0, \quad \varkappa_3^\pm(r)\geqslant 0.$$

Au cas où $q\equiv 0$ et $\varkappa_\alpha^\pm\equiv 0$ dans les conditions aux limites (3), (5), (7), (9) ou bien $l_1 = 0$ et sont données les conditions aux limites de deuxième espèce (5), (7), (9), on exige la satisfaction de la condition de résolubilité (voir (18), § 1).

Examinons toutes les variantes de combinaisons des conditions aux limites (2)-(9). Construisons les schémas aux différences correspondant aux conditions aux limites impliquées.

Introduisons dans le domaine $\bar{G}$ le maillage rectangulaire irrégulier quelconque

$$\bar{\omega} = \{(r_1, z_k) \in \bar{G},\ r_i = r_{i-1} + h_1(i),\ 1 \leqslant i \leqslant N_1,\ r_0 = l_1,\ r_{N_1} = L_1,\ z_k = z_{k-1} + h_3(k),\ 1 \leqslant k \leqslant N_3,\ z_0 = l_3,\ z_{N_3} = L_3\},$$

déterminons les pas moyens

$$\hbar_\alpha(m) = \begin{cases} 0{,}5h_\alpha(1), & m=0, \\ 0{,}5\,[h_\alpha(m) + h_\alpha(m+1), & 1 \leqslant m \leqslant N_\alpha - 1, \\ 0{,}5h_\alpha(N_\alpha), & m = N_\alpha, \quad \alpha = 1, 3, \end{cases}$$

et la fonction de maille d'une variable

$$\rho(i) = r_i, \quad 1 \leqslant i \leqslant N_1, \quad \rho(0) = \begin{cases} \frac{1}{4} h_1(1), & l_1 = 0, \\ l_1, & l_1 > 0. \end{cases}$$

Dans le cas primitif des coefficients continus k_1, k_3, q et f les coefficients du schéma aux différences seront définis par les formules

$$a_1(i, k) = \bar{r}_i k_1(\bar{r}_i, z_k), \quad a_3(i, k) = k_3(r_i, \bar{z}_k),$$
$$d(i, k) = q(r_i, z_k), \quad \varphi(i, k) = f(r_i, z_k),$$

où $\bar{r}_i = r_i - 0{,}5h_1(i)$, $\bar{z}_k = z_k - 0{,}5h_3(k)$.

En utilisant les notations introduites, approximons (1) aux équations aux différences

$$\frac{1}{\rho}(a_1 y_{\bar{r}})_{\hat{r}} + (a_3 y_{\bar{z}})_{\hat{z}} - dy = -\varphi, \quad 1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant k \leqslant N_3 - 1. \tag{10}$$

Les conditions aux limites de première espèce (2), (4), (6), (8) sont approximées de façon stricte:

$$y(0, k) = g_1^-(z_k), \quad 0 \leqslant k \leqslant N_3, \tag{11}$$
$$y(N_1, k) = g_1^+(z_k), \quad 0 \leqslant k \leqslant N_3, \tag{12}$$
$$y(i, 0) = g_3^-(r_i), \quad 0 \leqslant i \leqslant N_1, \tag{13}$$
$$y(i, N_3) = g_3^+(r_i), \quad 0 \leqslant i \leqslant N_1. \tag{14}$$

L'analogue au sens des différences finies des conditions aux limites (3) est de la forme

$$\frac{a_1^{+1}}{\rho\hbar_1} y_r + (a_3 y_{\bar{z}})_{\hat{z}} - \left(d + \frac{\varkappa_1^-}{\hbar_1}\right) y = -\varphi - \frac{g_1^-}{\hbar_1}, \quad i = 0, \tag{15}$$

où $1 \leqslant k \leqslant N_3 - 1$ et $\varkappa_1^- = g_1^- = 0$ si $l_1 = 0$. Les conditions aux limites (5), (7), (9) sont approximées de la façon suivante:

$$-\frac{a_1}{\rho\hbar_1} y_{\bar{r}} + (a_3 y_{\bar{z}})_{\hat{z}} - \left(d + \frac{\varkappa_1^+}{\hbar_1}\right) y = -\varphi - \frac{g_1^+}{\hbar_1}, \quad i = N_1, \tag{16}$$

où $1 \leqslant k \leqslant N_3 - 1$,

$$\frac{1}{\rho}(a_1 y_{\bar{r}})_{\hat{r}} + \frac{a_3^{+1}}{\hbar_3} y_z - \left(d + \frac{\varkappa_3^-}{\hbar_3}\right) y = -\varphi - \frac{g_3^-}{\hbar_3}, \quad k = 0, \tag{17}$$

$$\frac{1}{\rho}(a_1 y_{\bar{r}})_{\hat{r}} - \frac{a_3}{\hbar_3} y_{\bar{z}} - \left(d + \frac{\varkappa_3^+}{\hbar_3}\right) y = -\varphi - \frac{g_3^+}{\hbar_3}, \quad k = N_3, \tag{18}$$

où $1 \leqslant i \leqslant N_1 - 1$. On s'est servi des notations $a_1^{+1} = a_1(i+1, k)$, $a_3^{+1} = a_3(i, k+1)$.

Si aux côtés adjacents du rectangle $\bar{G}$ on impose les conditions aux limites de troisième espèce, aux nœuds d'angles du maillage $\bar{\omega}$ sont alors imposées les conditions aux limites

$$\frac{a_1^{+1}}{\rho \hbar_1} y_r + \frac{a_3^{+1}}{\hbar_3} y_z - \left(d + \frac{\varkappa_1^-}{\hbar_1} + \frac{\varkappa_3^-}{\hbar_3}\right) y = \\ = -\varphi - \frac{g_1^-}{\hbar_1} - \frac{g_3^-}{\hbar_3}, \quad i = k = 0, \tag{19}$$

$$-\frac{a_1}{\rho \hbar_1} y_{\bar{r}} + \frac{a_3^{+1}}{\hbar_3} y_z - \left(d + \frac{\varkappa_1^+}{\hbar_1} + \frac{\varkappa_3^-}{\hbar_3}\right) y = \\ = -\varphi - \frac{g_1^+}{\hbar_1} - \frac{g_3^-}{\hbar_3}, \quad i = N_1, \quad k = 0, \tag{20}$$

$$\frac{a_1^{+1}}{\rho \hbar_1} y_r - \frac{a_3}{\hbar_3} y_{\bar{z}} - \left(d + \frac{\varkappa_1^-}{\hbar_1} + \frac{\varkappa_3^+}{\hbar_3}\right) y = -\varphi - \frac{g_1^-}{\hbar_1} - \frac{g_3^+}{\hbar_3}, \\ i = 0, \quad k = N_3, \tag{21}$$

$$-\frac{a_1}{\rho \hbar_1} y_{\bar{r}} - \frac{a_3}{\hbar_3} y_{\bar{z}} - \left(d + \frac{\varkappa_1^+}{\hbar_1} + \frac{\varkappa_3^+}{\hbar_3}\right) y = -\varphi - \frac{g_1^+}{\hbar_1} - \frac{g_3^+}{\hbar_3}, \\ i = N_1, \quad k = N_3. \tag{22}$$

Comme auparavant, si $l_1 = 0$, il faut poser dans (19) et (21) $\varkappa_1^- = g_1^- = 0$.

Remarquons que le problème de différences (10), (15)-(22) avec conditions aux limites de troisième espèce sur chacun des côtés du rectangle $\bar{G}$ peut être écrit sous forme compacte

$$\Lambda y = -f, \quad 0 \leqslant i \leqslant N_1, \quad 0 \leqslant k \leqslant N_3, \tag{23}$$
$$\Lambda = \Lambda_1 + \Lambda_3, \quad f = \varphi + \varphi_1/\hbar_1 + \varphi_3/\hbar_3,$$

où

$$\varphi_1(i,k) = \begin{cases} g_1^-, & i = 0, \\ 0, & 1 \leqslant i \leqslant N_1 - 1, \\ g_1^+, & i = N_1, \end{cases}$$
$$\varphi_3(i,k) = \begin{cases} g_3^-, & k = 0, \\ 0, & 1 \leqslant k \leqslant N_3 - 1, \\ g_3^+, & k = N_3, \end{cases} \tag{24}$$

tandis que les opérateurs de différences Λ_1 et Λ_3 sont définis par les formules

$$\Lambda_1 y = \begin{cases} \frac{a_1^{+1}}{\rho \hbar_1} y_r - \left(d_1 + \frac{\varkappa_1^-}{\hbar_1}\right) y, & i = 0, \\ \frac{1}{\rho} (a_1 y_{\bar{r}})_{\hat{r}} - d_1 y, & 1 \leqslant i \leqslant N_1 - 1, \\ -\frac{a_1}{\rho \hbar_1} y_{\bar{r}} - \left(d_1 + \frac{\varkappa_1^+}{\hbar_1}\right) y, & i = N_1, \quad 0 \leqslant k \leqslant N_3, \end{cases} \tag{25}$$

$$\Lambda_3 y = \begin{cases} \frac{a_3^{+1}}{\hbar_3} y_z - \left(d_3 + \frac{\varkappa_3^-}{\hbar_3}\right) y, & k = 0, \\ (a_3 y_{\bar{z}})_{\hat{z}} - d_3 y, & 1 \leqslant k \leqslant N_3 - 1, \\ -\frac{a_3}{\hbar_3} y_{\bar{z}} - \left(d_3 + \frac{\varkappa_3^+}{\hbar_3}\right) y, & k = N_3, \quad 0 \leqslant i \leqslant N_1. \end{cases} \tag{26}$$

Ici $d_1 + d_3 = d$, $d_1 \geqslant 0$ et $d_3 \geqslant 0$.

Cherchons les conditions de la résolubilité du schéma aux différences (23) au cas où $d \equiv 0$ et $\varkappa_\alpha^\pm \equiv 0$, $\alpha = 1, 2$.

Dans l'espace H des fonctions de mailles associées à $\bar{\omega}$ définissons le produit scalaire suivant la formule

$$(u, v) = \sum_{i=0}^{N_1} \sum_{k=0}^{N_3} u(i, k) v(i, k) \rho(i) \hbar_1(i) \hbar_3(k). \tag{27}$$

Déterminons les opérateurs A_1 et A_3 agissant dans H en posant $A_\alpha = -\Lambda_\alpha$, $\alpha = 1, 3$. Le schéma aux différences (23) peut alors s'écrire sous la forme d'une équation opératorielle

$$Au = f, \quad A = A_1 + A_3. \tag{28}$$

En utilisant la première formule de différences de Green on obtient pour le cas de $d \equiv 0$ et $\varkappa_\alpha^\pm \equiv 0$ que

$$(Au, v) = \sum_{i=1}^{N_1} \sum_{k=0}^{N_3} h_1(i) \hbar_3(k) a_1 u_{\bar{r}} v_{\bar{r}})_{ik} + \\ + \sum_{i=0}^{N_1} \sum_{k=1}^{N_3} \hbar_1(i) h_3(k) \rho(i) (a_3 u_{\bar{z}} v_{\bar{z}})_{ik} = (u, Av).$$

Par conséquent, l'opérateur A est autoadjoint dans H et non négatif, avec $(Au, u) = 0$ seulement dans le cas où $u(i, k) \equiv \text{const}$ ou $u(i, k) \equiv 0$. De là, en vertu de l'inégalité de Cauchy-Bouniakovski

$$(Au, u^2) \leqslant (Au, Au)(u, u),$$

il résulte que $Au = 0$ pour $u \neq 0$, si u est une constante sur $\bar{\omega}$. Le noyau de l'opérateur A est donc constitué de fonctions de mailles égales à des constantes sur le maillage $\bar{\omega}$. Le problème (28) est donc

résoluble si la condition $(f, 1) = 0$ est satisfaite, ou bien si, en vertu de la définition de f, la condition

$$\sum_{i=0}^{N_1} \sum_{k=0}^{N_3} \rho\varphi \hbar_1 \hbar_2 + \sum_{k=0}^{N_3} \hbar_3 (\rho g_1^- + \rho g_1^+) + \sum_{i=0}^{N_1} \hbar_i \rho [g_3^- + g_3^+] = 0 \qquad (29)$$

est remplie.

La condition (29) est un analogue au sens des différences finies de la condition (18) de résolubilité du problème différentiel correspondant au problème de différences (23).

Si la condition (29) est remplie, la solution du problème (23) existe pour $d \equiv 0$ et $\varkappa_\alpha^\pm \equiv 0$, toutefois elle n'est pas unique et deux solutions quelconques diffèrent d'une constante. Une des solutions peut donc être séparée en fixant la valeur de $y\ (i, k)$ en l'un quelconque des nœuds du maillage $\bar{\omega}$.

2. Méthodes directes. Considérons le cas pour lequel les problèmes de différences (10)-(22) peuvent être résolus par l'une des méthodes directes exposées dans les chapitres III et IV.

Supposons que *les coefficients* k_1, k_3 et q *de l'équation* (1) *ne dépendent pas* de z, c'est-à-dire que $k = k_1(r)$, $k_3 = k_3(r)$, $q = q(r)$, dans les conditions aux limites de troisième espèce (3), (5) les coefficients $\varkappa_1^+$ et $\varkappa_1^-$ sont constants, tandis que dans les conditions (7), (9) $\varkappa_3^- = \varkappa_3^+ \equiv 0$.

Toutes combinaisons des conditions aux limites (2)-(9) sont possibles. On admet que le maillage $\bar{\omega}$ est régulier en z, c'est-à-dire que $h_3(k) \equiv h_3$ et peut être irrégulier en r. Avec ces hypothèses les problèmes de différences (10)-(22) peuvent être résolus soit par la méthode de réduction totale, soit par la méthode combinée de réduction incomplète et de séparation des variables.

Illustrons la possibilité d'application des méthodes directes par un exemple où aux côtés $r = l_1$ et $r = L_1$ sont imposées des conditions aux limites de troisième (ou deuxième) espèce (3), (5), pour $z = l_3$ et $z = L_3$ de deuxième espèce. Les autres combinaisons des conditions aux limites sont étudiées de façon analogue.

Le schéma aux différences correspondant au problème posé est de la forme (23). En vertu des hypothèses ci-dessus émises, les coefficients du schéma aux différences se déterminent suivant les formules (comp. avec point 1) $a_1 = a_1(i) = \bar{r}_i k_1(\bar{r}_i)$, $a_3 = a_3(i) = k_3(r_i)$, $d = d(i) = q(r_i)$, de sorte que $a_3^{+1} = a_3$. Dans la définition (25) de l'opérateur de différences Λ_1 choisissons $d_1 = d$, tandis que dans les formules (26) donnant l'opérateur Λ_3 posons $\varkappa_3^- = \varkappa_3^+ = 0$, $d_3 = 0$. Vu que le maillage $\bar{\omega}$ est régulier en z, il faut remplacer dans (26) l'expression de différences $(a_3 y_{\bar{z}})_{\hat{z}}$ par l'expression $a_3 y_{\bar{z}z}$.

Réduisons à présent le problème de différences (23) à un système d'équations vectorielles triponctuelles. A cet effet introduisons le vecteur d'inconnues

$$Y_k = (y(0, k), y(1, k), \ldots, y(N_1, k)), \quad 0 \leqslant k \leqslant N_3,$$

contenant la valeur de la fonction de maille cherchée sur la k-ième ligne du maillage $\bar{\omega}$ et le vecteur des seconds membres

$$F_k = (\theta_0 f(0, k), \theta_1 f(1, k), \ldots, \theta_{N_1} f(N_1, k)), \qquad 0 \leqslant k \leqslant N_3,$$

où $\theta_i = h_3^2/a_3(i)$, $0 \leqslant i \leqslant N_1$. Définissons la matrice carrée C en posant

$$CY_k = ((2E - \theta_0\Lambda_1) y(0, k), \ldots, (2E - \theta_{N_1}\Lambda_1) y(N_1, k)).$$

En se servant de ces notations, écrivons le schéma aux différences (23) sous forme vectorielle

$$CY_0 - 2Y_1 = F_0, \qquad k = 0,$$

$$-Y_{k-1} + CY_k - Y_{k+1} = F_k, \qquad 1 \leqslant k \leqslant N_3 - 1, \qquad (30)$$

$$2Y_{N_3-1} + CY_{N_3} = F_{N_3}, \qquad k = N_3.$$

Pour s'en convaincre, il suffit de multiplier chaque équation du schéma (23) par $(-\theta_i)$ et de passer à l'écriture vectorielle.

Rappelons que la méthode de réduction totale a été construite pour le système (30) au point 1, § 4, ch. III. La méthode combinée de réduction incomplète et de séparation des variables a été construite au point 2, § 3, ch. IV. Dans le cas concerné, à la différence des exemples examinés aux chapitres III et IV, l'opérateur Λ_1 est défini d'une autre manière. Mais puisque l'opérateur Λ_1 est toujours triponctuel, la différence observée n'exerce aucune influence sur la construction de ces méthodes, ainsi que sur la nature des rapports entre le nombre d'opérations arithmétiques et celui des nœuds du maillage $\bar{\omega}$. Si $N_3 = 2^n$, le nombre d'opérations arithmétiques des méthodes concernées s'apprécie par la quantité $O(N_1 N_3 \log_2 N_3)$.

Notons en conclusion que l'application de la méthode combinée avec séparation de l'une des solutions dans le cas dégénéré ($d \equiv 0$, $\varkappa_1^- = \varkappa_2^+ \equiv 0$) est décrite en détail au point 2, § 4, ch. XII pour le système des coordonnées cartésiennes.

3. Méthode des directions alternées. Examinons maintenant le cas particulier du problème (1)-(9) pour lequel $k_1 = k_1(r)$, $k_3 = k_3(z)$, $q =$ const, $\varkappa_\alpha^\pm =$ const, $\alpha = 1, 3$, tandis qu'aux côtés du rectangle $\bar{G}$ est imposée une combinaison quelconque des conditions aux limites (2)-(9). Dans ce cas les variables du problème (1)-(9) se divisent.

Il est admis que le maillage $\bar{\omega}$ est quelconque et irrégulier suivant chaque direction. Avec les hypothèses émises, les problèmes de différences (10)-(22) peuvent être résolus par la méthode des directions alternées avec le jeu optimal de paramètres d'itération donnée au chapitre XI pour le cas d'un système de coordonnées cartésiennes.

Illustrons l'application de cette méthode par un exemple dans lequel aux côtés du rectangle $\bar{G}$ sont imposées des conditions aux limites de troisième espèce (3), (5), (7), (9). Le schéma aux différences correspondant au problème (1), (3), (5), (7), (9) est de la forme (23), où les opérateurs Λ_1 et Λ_3 sont définis dans (25) (26), tandis que les coefficients a_1, a_3, d_1 et d_3 sont donnés par les formules $a_1(i) = \bar{r}_i k_1(\bar{r}_i)$, $a_3(k) = k_3(\bar{z}_k)$, $d_1 = d_3 = 0{,}5d$, $d = q$.

Au point 1 on a montré que le problème de différences (23) peut être écrit sous forme de l'équation opératorielle (28)

$$Au = f, \quad A = A_1 + A_3$$

dans l'espace hilbertien H des fonctions de mailles associées à $\bar{\omega}$. Indiquons les propriétés principales des opérateurs A_1 et A_3:

1) les opérateurs A_1 et A_3 sont permutables, $A_1A_3 = A_3A_1$;

2) A_1 et A_3 sont des opérateurs autoadjoints, $(A_\alpha u, v) = (u, A_\alpha v)$;

3) les opérateurs A_1 et A_3 sont des opérateurs non négatifs bornés, c'est-à-dire que pour tout $u \in H$ sont satisfaites les inégalités

$$\delta_\alpha (u, u) \leqslant (A_\alpha u, u) \leqslant \Delta_\alpha (u, u), \qquad \delta_\alpha \geqslant 0, \ \Delta_\alpha > 0, \ \alpha = 1, 3. \tag{31}$$

En effet, la permutabilité des opérateurs A_1 et A_3 s'ensuit de la structure des opérateurs A_1 et A_3 et de l'hypothèse relativement aux coefficients k_1, k_3, q et $\varkappa_\alpha^\pm$.

Ensuite, en utilisant la définition (27) du produit scalaire dans H et les formules de différences de Green, on aboutit pour A_1 et $u, v \in H$ quelconques à l'égalité

$$(A_1u, v) = \sum_{i=1}^{N_1} \sum_{k=0}^{N_3} h_1(i)\, \hbar_3(k)\, (a_1 u_{\bar{r}} v_{\bar{r}})_{ik} + d_1(u, v) + \sum_{k=0}^{N_3} \hbar_3(k) \left[\varkappa_1^- \rho uv|_{i=0} + \varkappa_1^+ \rho uv|_{i=N_1}\right] \tag{32}$$

et à une égalité analogue pour A_3

$$(A_3u, v) = \sum_{k=1}^{N_3} \sum_{i=0}^{N_1} \rho(i)\, \hbar_1(i)\, h_3(k)\, (a_3 u_{\bar{z}} v_{\bar{z}})_{ik} + d_3(u, v) + \sum_{i=0}^{N_1} \rho(i)\, \hbar_1(i) \left[\varkappa_3^- uv|_{k=0} + \varkappa_3^+ uv|_{k=N_3}\right]. \tag{33}$$

En changeant de place u et v on se convainc que les opérateurs A_1 et A_3 sont autoadjoints.

Si l'on pose ici $u = v$ et l'on tient compte de la condition $k_1 \geqslant \geqslant c_1 > 0$, $k_3 \geqslant c_1 > 0$, $q \geqslant 0$, $\varkappa_\alpha^\pm \geqslant 0$, $\alpha = 1, 3$, on constatera que les opérateurs A_1 et A_3 sont non négatifs, c'est-à-dire que $(A_\alpha u, u) \geqslant 0$. Si est satisfaite la condition

$$d_\alpha^2 + (\varkappa_\alpha^-)^2 + (\varkappa_\alpha^+)^2 \neq 0, \quad \alpha = 1, 3, \tag{34}$$

δ_α est alors positif. Admettons que (34) est satisfait.

Apprécions δ_α par le bas.

A partir du lemme 16 du chapitre V on obtient pour un i fixé, $0 \leqslant i \leqslant N_1$ l'estimation

$$\delta_3 \sum_{k=0}^{N_3} \hbar_3(k)\, u^2(i, k) \leqslant \sum_{k=1}^{N_3} h_3(k)\, a_3(k)\, u_{\bar z}^2(i, k) + \\ + d_3 \sum_{k=0}^{N_3} \hbar_3(k)\, u^2(i, k) + \varkappa_3^- u^2(i, 0) + \varkappa_3^+ u^2(i, N_3), \tag{35}$$

où $1/\delta_3 = \max\limits_{0 \leqslant k \leqslant N_3} v(k)$, $v(k)$ étant la solution du problème aux limites

$$\begin{aligned} &(a_3 v_{\bar z})_{\hat z} - d_3 v = -1, \quad 1 \leqslant k \leqslant N_3 - 1, \\ &\frac{a_3^{+1}}{\hbar_3} v_z - \left(d_3 + \frac{\varkappa_3^-}{\hbar_3}\right) v = -1, \quad k = 0, \\ &-\frac{a_3}{\hbar_3} v_{\bar z} - \left(d_3 + \frac{\varkappa_3^+}{\hbar_3}\right) v = -1, \quad k = N_3. \end{aligned} \tag{36}$$

Comme la condition (34) est remplie, la solution du problème (36) existe et est unique. En multipliant maintenant (35) par $\rho(i)\, \hbar_1(i)$ et en sommant en i de 0 à N_1, on obtient l'inégalité $\delta_3 (u, u) \leqslant (A_3 u, u)$. En résolvant numériquement le problème (36), on détermine δ_3. Bref, on a trouvé la constante δ_3. De façon analogue est appréciée la constante δ_1: $1/\delta_1 = \max\limits_{0 \leqslant i \leqslant N_1} \bar v(i)$, où $\bar v(i)$ est la solution du problème aux limites

$$\begin{aligned} &\frac{1}{\rho} (a_1 \bar v_{\bar r})_{\hat r} - d_1 \bar v = -1, \quad 1 \leqslant i \leqslant N_1 - 1, \\ &\frac{a_1^{+1}}{\rho \hbar_1} \bar v_r - \left(d_1 + \frac{\varkappa_1^-}{\hbar_1}\right) \bar v = -1, \quad i = 0, \\ &-\frac{a_1}{\rho \hbar_1} \bar v_{\bar r} - \left(d_1 + \frac{\varkappa_1^+}{\hbar_1}\right) \bar v = -1, \quad i = N_1. \end{aligned} \tag{37}$$

Cherchons maintenant les estimations pour Δ_1 et Δ_3. A partir de (33) pour $u = v$, on obtient

$$(A_3 u, u) = \sum_{i=0}^{N_1} \rho(i)\, \hbar_1(i) \Big[\sum_{k=1}^{N_3} h_3(k)\, a_3(k)\, u_{\bar z}^2(i, k) + \\ + d_3 \sum_{k=0}^{N_3} \hbar_3(k)\, u^2(i, k) + \varkappa_3^- u^2(i, 0) + \varkappa_3^+ u^2(i, N_3) \Big].$$

Apprécions l'expression entre les crochets. A partir du lemme 16, ch. V il se dégage

$$d_3 \sum_{k=0}^{N_3} u^2(i, k)\,\hbar_3(k) + \varkappa_3^- u^2(i, 0) + \varkappa_3^+ u^2(i, N_3) \leqslant$$

$$\leqslant m_1 \Big[\sum_{k=1}^{N_3} a_3(k)\, u_{\bar z}^2(i, k)\, h_3(k) + \sum_{k=0}^{N_3} \hbar_3(k)\, u^2(i, k) \Big], \qquad (38)$$

où $m_1 = \max\limits_{0 \leqslant k \leqslant N_3} w(k)$, tandis que $w(k)$ est la solution du problème aux limites

$$\begin{aligned} &(a_3 w_{\bar z})_{\hat z} - w = -d_3, \quad 1 \leqslant k \leqslant N_3 - 1, \\ &\frac{a_3^{+1}}{\hbar_3} w_z - w = -\Big(d_3 + \frac{\varkappa_3^-}{\hbar_3}\Big), \quad k = 0, \\ &-\frac{a_3}{\hbar_3} w_{\bar z} - w = -\Big(d_3 + \frac{\varkappa_3^+}{\hbar_3}\Big), \quad k = N_3. \end{aligned} \qquad (39)$$

En utilisant le lemme 17, ch. V, on aura

$$\sum_{k=1}^{N_3} a_3(k)\, u_{\bar z}^2(i, k)\, h_3(k) \leqslant m_2 \sum_{k=0}^{N_3} \hbar_3(k)\, u^2(i, k), \qquad (40)$$

où

$$m_2 = \max \left(\frac{a_3(N_3)}{\hbar_3^2(N_3)}, \frac{a_3(1)}{\hbar_3^2(0)}, \max_{1 \leqslant k \leqslant N_3 - 1} \frac{2}{\hbar_3(k)} \Big[\frac{a_3(k)}{h_3(k)} + \frac{a_3(k+1)}{h_3(k+1)} \Big] \right).$$

Il s'ensuit de (38) et (40) l'estimation

$$\sum_{k=1}^{N_3} h_3(k)\, a_3(k)\, u_{\bar z}^2(i, k) + d_3 \sum_{k=0}^{N_3} \hbar_3(k)\, u^2(i, k) + \varkappa_3^- u^2(i, 0) +$$

$$+ \varkappa_3^+ u^2(i, N_3) \leqslant \Delta_3 \sum_{k=0}^{N_3} \hbar_3(k)\, u^2(i, k), \qquad \Delta_3 = m_1 + m_2(1 + m_1).$$

En multipliant l'inégalité obtenue par $\rho(i)\,\hbar_1(i)$ et en la sommant en i de 0 à N_1 on aura l'estimation $(A_3 u, u) \leqslant \Delta_3 (u, u)$.

De façon analogue on trouve Δ_1: $\Delta_1 = \bar m_1 + \bar m_2 (1 + \bar m_1)$, où $\bar m_1 = \max\limits_{0 \leqslant i \leqslant N_1} \bar w(i)$, $\bar w(i)$ étant la solution du problème aux limites

$$\begin{aligned} &\frac{1}{\rho} (a_1 \bar w_{\bar r})_{\hat r} - \bar w = -d_1, \quad 1 \leqslant i \leqslant N_1 - 1, \\ &\frac{a_1^{+1}}{\rho \hbar_1} \bar w_r - \bar w = -\Big(d_1 + \frac{\varkappa_1^-}{\hbar_1}\Big), \quad i = 0, \\ &-\frac{a_1}{\rho \hbar_1} \bar w_{\bar r} - \bar w = -\Big(d_1 + \frac{\varkappa_1^+}{\hbar_1}\Big), \quad i = N_1, \end{aligned} \qquad (41)$$

avec

$$\bar{m}_1 = \max\left(\frac{a_1(N_1)}{\rho(N_1)\hbar_1^2(N_1)},\ \frac{a_1(1)}{\rho(0)\hbar_1^2(0)},\ \max_{1\leqslant i\leqslant N_1-1}\frac{2}{\rho(i)\hbar_1(i)}\left[\frac{a_1(i)}{h_1(i)}+\frac{a_1(i+1)}{h_1(i+1)}\right]\right).$$

En résolvant numériquement le problème (41) on détermine $\bar{m}_1$ et, partant, Δ_1. On a ainsi trouvé les constantes δ_α et Δ_α, $\alpha = 1, 3$, figurant dans les inégalités (31).

Rappelons que le schéma itératif de la méthode des directions alternées appliquée à l'équation opératorielle (28) est de la forme (voir ch. XI)

$$B_{k+1}\frac{y_{k+1}-y_k}{\tau_{k+1}}+Ay_k=f,\quad k=0,1,\ldots,\quad y_0\in H,$$
$$B_k=(\omega_k^{(1)}E+A_1)(\omega_k^{(3)}E+A_3),\quad \tau_k=\omega_k^{(1)}+\omega_k^{(3)}. \tag{42}$$

Au point 4, § 1, ch. XI on a construit pour le schéma itératif (42), dont les opérateurs A_1 et A_3 satisfont aux propriétés 1)-3) susmentionnées, le jeu optimal des paramètres $\omega_k^{(1)}$ et $\omega_k^{(2)}$, $k = 1, 2, \ldots, n$. En utilisant ce jeu de paramètres, la précision relative $\varepsilon > 0$ ($\| y_n - u \|_D \leqslant \varepsilon \| y_0 - u \|_D$, $D = A, E$) est atteinte si l'on effectue $n \geqslant n_0(\varepsilon)$ itérations, où

$$n_0(\varepsilon)=\frac{1}{\pi^2}\ln\frac{4}{\eta}\ln\frac{4}{\varepsilon},\quad \eta=\frac{1-a}{1+a},\quad a=\sqrt{\frac{(\Delta_1-\delta_1)(\Delta_3-\delta_3)}{(\Delta_1+\delta_3)(\Delta_3+\delta_1)}}.$$

Le jeu des paramètres optimaux $\omega_k^{(1)}$ et $\omega_k^{(2)}$ pour le second problème aux limites ($d = 0$, $\varkappa_\alpha^\pm \equiv 0$) a été construit au point 1, § 4, ch. XII.

4. Résolution d'équations données sur la surface d'un cylindre. Voyons à présent la méthode de résolution des analogues au sens des différences finies des problèmes aux limites pour une équation elliptique sans dérivées mixtes, donnée sur la surface d'un cylindre de rayon R. Limitons-nous à l'examen d'une surface de cylindre fermée suivant φ, vu que les méthodes de résolution des problèmes au cas de surface non fermée ne diffèrent en rien de celles des problèmes plans avec variables cartésiennes.

Bref, on recherche dans le domaine $\bar{G} = \{l_2 \leqslant \varphi \leqslant L_2,\ l_3 \leqslant z \leqslant L_3,\ L_2 - l_2 = 2\pi\}$ la solution de l'équation

$$\frac{1}{R^2}\frac{\partial}{\partial\varphi}\left(k_2\frac{\partial u}{\partial\varphi}\right)+\frac{\partial}{\partial z}\left(k_3\frac{\partial u}{\partial z}\right)-qu=-f(\varphi,z),\quad (\varphi,z)\in G, \tag{43}$$

périodique en φ de période 2π satisfaisant sur les côtés $z = l_3$ et $z = L_3$ soit aux conditions aux limites de première espèce $u(\varphi, z) = g_3^-(\varphi)$ pour $z = l_3$, $u(\varphi, z) = g_3^+(\varphi)$ pour $z = L_3$, soit de deu-

xième ou de troisième espèce

$$\begin{aligned} k_3 \frac{\partial u}{\partial z} &= \varkappa_3^- u - g_3^-(\varphi), \quad z = l_3, \\ -k_3 \frac{\partial u}{\partial z} &= \varkappa_3^+ u - g_3^+(\varphi), \quad z = L_3, \end{aligned} \tag{44}$$

roit à leur combinaison quelconque. On admet que les coefficients semplissent les conditions

$$k_2(\varphi, z) \geqslant c_1 > 0, \quad k_3(\varphi, z) \geqslant c_1 > 0,$$
$$q(\varphi, z) \geqslant 0, \quad \varkappa_3^{\pm}(\varphi) \leqslant 0.$$

Introduisons dans le domaine $\bar{G}$ un maillage irrégulier quelconque

$$\bar{\omega} = \{(\varphi_j, z_k) \in \bar{G}, \ \varphi_j = \varphi_{j-1} + h_2(j), \ 1 \leqslant j \leqslant N_2, \ \varphi_0 = l_2,$$
$$\varphi_{N_2} = L_2, \ z_k = z_{k-1} + h_3(k),$$
$$1 \leqslant k \leqslant N_3, \ z_0 = l_3, \ z_{N_3} = L_3\}$$

et définissons le pas moyen

$$\hbar_2(j) = \begin{cases} 0{,}5\,[h_2(1) + h_2(N_2)], & j = 0, \\ 0{,}5\,[h_2(j) + h_2(j+1)], & 1 \leqslant j \leqslant N_2 - 1. \end{cases} \tag{45}$$

Le pas moyen $\hbar_3(k)$ a été défini plus haut.

L'équation (43), compte tenu de sa périodicité, s'approxime de la façon suivante:

$$(a_2 y_{\bar{\varphi}})_{\hat{\varphi}} + (a_3 y_{\bar{z}})_{\hat{z}} - dy = -\psi, \quad 0 \leqslant j \leqslant N_2 - 1, \quad 1 \leqslant k \leqslant N_3 - 1, \tag{46}$$

où l'on utilise les relations $y(j, k) = y(N_2 + j, k)$, $j = 0, -1$, $a_2(0, k) = a_2(N_2, k)$, $h_2(0) = h_2(N_2)$, qui sont des corollaires de la périodicité. Au cas de coefficients lisses k_2, k_3, q et f les coefficients de l'équation (46) peuvent, par exemple, être choisis de la sorte:

$$a_2(j, k) = \frac{1}{R^2} k_2(\varphi_j - 0{,}5h_2(j), z_k), \quad d(j, k) = q(\varphi_j, z_k),$$
$$a_3(j, k) = k_3(\varphi_j, z_k - 0{,}5h_3(k)), \quad \psi(j, k) = f(\varphi_j, z_k).$$

Les conditions aux limites de première espèce s'approximent de façon exacte

$$y(j, 0) = g_3^-(\varphi_j), \ k = 0, \ y(j, N_3) = g_3^+(\varphi_j), \quad k = N_3 \tag{47}$$

pour $0 \leqslant j \leqslant N_2 - 1$, tandis que l'analogue au sens des différences finies des conditions aux limites (44) de troisième espèce prend pour $0 \leqslant j \leqslant N_2 - 1$ la forme

$$\begin{aligned} (a_2 y_{\bar{\varphi}})_{\hat{\varphi}} + \frac{a_3^{+1}}{\hbar_3} y_z - \left(d + \frac{\varkappa_3^-}{\hbar_3}\right) y &= -\psi - \frac{g_3^-}{\hbar_3}, \quad k = 0, \\ (a_2 y_{\bar{\varphi}})_{\hat{\varphi}} - \frac{a_3}{\hbar_3} y_{\bar{z}} - \left(d + \frac{\varkappa_3^+}{\hbar_3}\right) y &= -\psi - \frac{g_3^+}{\hbar_3}, \quad k = N_3. \end{aligned} \tag{48}$$

Dans le problème (46), (47) les inconnues sont les valeurs $y(j, k)$ pour $0 \leqslant j \leqslant N_2 - 1$, $1 \leqslant k \leqslant N_3 - 1$, tandis que dans le problème (46), (48) elles s'obtiennent pour les mêmes valeurs de j et pour $0 \leqslant k \leqslant N_3$.

Cherchons les conditions de résolubilité du problème de différences (46), (48) pour le cas où $d \equiv 0$, $\varkappa_3^{\pm} \equiv 0$. Écrivons d'abord le schéma (46), (48) sous la forme

$$\Lambda y = -f, \quad 0 \leqslant j \leqslant N_2 - 1, \quad 0 \leqslant k \leqslant N_3,$$
$$\Lambda = \Lambda_2 + \Lambda_3, \quad f = \psi + \psi_3/\hbar_3, \tag{49}$$

où l'opérateur de différences Λ_3 est défini dans (26) avec $d_3 = d$, et l'opérateur Λ_2 est donné par la formule $\Lambda_2 y = (a_2 y_{\bar{\varphi}})_{\hat{\varphi}}$, $0 \leqslant j \leqslant N_2 - 1$,

$$\psi_3(j, k) = \begin{cases} g_3^-(\varphi_j), & k = 0, \\ 0, & 1 \leqslant k \leqslant N_3 - 1, \\ g_3^+(\varphi_j), & k = N_3. \end{cases}$$

Soient maintenant $d \equiv 0$ et $\varkappa_3^{\pm} \equiv 0$. Désignons par H l'espace des fonctions de mailles associées à $\bar{\omega}^* = \{(\varphi_j, z_k) \in \bar{\omega},\ 0 \leqslant j \leqslant N_2 - 1,\ 0 \leqslant k \leqslant N_3\}$ dont le produit scalaire sera déterminé par la formule

$$(u, v) = \sum_{j=0}^{N_2-1} \sum_{k=0}^{N_3} u(j, k)\, v(j, k)\, \hbar_2(j)\, \hbar_3(k).$$

Définissons les opérateurs A_2 et A_3 agissant dans H par les égalités: $A_3 = -\Lambda_3$, $A_2 y = -\Lambda_2 \bar{y}$, où $y(j, k) = \bar{y}(j, k)$ pour $0 \leqslant j \leqslant N_2 - 1$, $0 \leqslant k \leqslant N_3$ et $\bar{y}$ satisfait à la condition de périodicité $\bar{y}(j, k) = \bar{y}(N_2 + j, k)$, $j = 0, -1$.

En utilisant les notations introduites, écrivons le schéma aux différences (49) sous forme d'une équation opératorielle

$$Au = f, \quad A = A_2 + A_3. \tag{50}$$

Compte tenu des conditions de périodicité on obtient à l'aide de la formule de différences de Green

$$(Au, v) = -(\Lambda \bar{u}, \bar{v}) = \sum_{k=0}^{N_3} \sum_{j=0}^{N_2-1} \hbar_3(k)\, h_2(j)\, (a_2 \bar{u}_{\bar{\varphi}} \bar{v}_{\bar{\varphi}})_{jk} + \sum_{k=1}^{N_3} \sum_{j=0}^{N_2-1} \hbar_2(j)\, h_3(k)\, (a_3 \bar{u}_{\bar{z}} \bar{v}_{\bar{z}})_{jk} = (u, Av).$$

Par conséquent, l'opérateur A est autoadjoint dans H. En outre, en examinant les valeurs de (Au, u), on constate que le noyau de l'opérateur A est composé de fonctions de mailles qui prennent sur le mail-

lage ω^* des valeurs constantes. La solution du problème de différences (49) existe donc si la condition $(f, 1) = 0$ est remplie. En y portant f de (49), on obtient

$$\sum_{j=0}^{N_2-1} \sum_{k=0}^{N_3} \hbar_2(j)\, \hbar_3(k)\, \psi(j, k) + \sum_{j=0}^{N_2-1} \hbar_2(j)\, [g_3^-(\varphi_j) + g_3^+(\varphi_j)] = 0.$$

Avec la satisfaction de cette condition la solution du problème de différences (46), (48) pour $d \equiv 0$ et $\varkappa_3^{\pm} = 0$ existe et deux quelconques de ses solutions diffèrent d'une constante.

Voyons les cas où la solution des problèmes de différences (46)-(48) peut être obtenue par des méthodes directes exposées aux chapitres III et IV.

P r e m i e r c a s. Les coefficients k_2, k_3 et q de l'équation (43) ne dépendent que de φ, $\varkappa_3^{\pm} = \text{const}$ et le maillage $\bar{\omega}$ est régulier en z. Le problème de différences (46), (48) peut être écrit sous forme d'un système d'équations vectorielles triponctuelles

$$\begin{aligned} (C + 2\alpha E)\, Y_0 - 2Y_1 &= F_0, & k &= 0, \\ -Y_{k-1} + CY_k - Y_{k+1} &= F_k, & 1 \leqslant k &\leqslant N_3 - 1, \\ -2Y_{N_3-1} + (C + 2\beta E)\, Y_{N_3} &= F_{N_3}, & k &= N_3, \end{aligned} \qquad (51)$$

où $N_3 = 2^n$, $n > 0$ est un nombre entier,

$$Y_k = (y(0, k), y(1, k), \ldots, y(N_2 - 1, k)),$$

$$F_k = (\theta_0 f(0, k), \theta_1 f(1, k), \ldots, \theta_{N_2-1} f(N_2 - 1, k)),$$

$$CY_k = ((2E - \theta_0\Lambda_2)\, y(0, k), \ldots, (2E - \theta_{N_2-1}\Lambda_2)\, y(N_2 - 1, k))$$

pour $0 \leqslant k \leqslant N_3$. L'opérateur Λ_2 est défini plus haut, $f(j, k)$ est donné dans (49) et $\theta_j = h_3^2/a_3(j)$, $\alpha = h_3\varkappa_3^-$, $\beta = h_3\varkappa_3^+$.

Rappelons qu'au point 3, § 4, ch. III, lors de la résolution du problème (51) avec la condition $\alpha^2 + \beta^2 \neq 0$, on avait construit la méthode de réduction totale. Si $\alpha = \beta = 0$, mais $d \not\equiv 0$ l'algorithme de la méthode est exposé au point 1, § 4, ch. III. Pour ce dernier cas on a construit au point 2, § 3, ch. IV la méthode combinée de réduction incomplète et de séparation des variables.

D e u x i è m e c a s. Les coefficients k_2, k_3 et q ne dépendent que de z, $\varkappa_3^{\pm} = \text{const}$ et le maillage $\bar{\omega}$ est régulier en φ. Le problème de différences (46), (48) s'écrit sous la forme d'un système d'équations vectorielles triponctuelles

$$\begin{aligned} -Y_{N_2-1} + CY_0 - Y_1 &= F_0 & j &= 0, \\ -Y_{j-1} + CY_j - Y_{j+1} &= F_j, & 1 \leqslant j &\leqslant N_2 - 2, \\ -Y_{N_2-2} + CY_{N_2-1} - Y_0 &= F_{N_2-1}, & j &= N_2 - 1. \end{aligned} \qquad (52)$$

N_2 est ici égal à 2^n, $n > 0$ étant un nombre entier,

$$Y_j = (y(j, 0), y(j, 1), \ldots, y(j, N_3)),$$

$$F_j = (\theta_0 f(j, 0), \theta_1 f(j, 1), \ldots, \theta_{N_3} f(j, N_3)),$$

$$CY_j = ((2E - \theta_0\Lambda_3)\, y(j, 0), \ldots, (2E - \theta_{N_3}\Lambda_3)\, y(j, N_3)),$$

où $0 \leqslant j \leqslant N_2 - 1$. L'opérateur de différences Λ_3 est défini dans (26), avec $d_3 = d$ et $\theta_k = h_3^2/a_2(k)$, $0 \leqslant k \leqslant N_3$. Le problème (52) peut être résolu en recourant à la méthode de réduction totale construite au point 2, § 4, ch. III ou au moyen de la méthode combinée, où est utilisé l'algorithme de transformation discrète rapide de Fourier d'une fonction périodique réelle. Cet algorithme est construit au point 4, § 1, ch. IV.

Dans chacun des cas passés en revue les méthodes directes sont mises en œuvre en $O(N_2N_3n)$ opérations.

En conclusion, notons que si les coefficients satisfont aux conditions $k_2 = k_2(\varphi)$, $k_3 = k_3(z)$, $g = \text{const}$, $\varkappa_3^\pm = \text{const}$, tandis que le maillage est irrégulier suivant chaque direction, alors pour la résolution du problème (46), (48) on peut recourir à la méthode des directions alternées avec un jeu optimal de paramètres:

$$B_{k+1}\frac{y_{k+1}-y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \ldots, \quad y_0 \in H,$$

$$B_k = (\omega_k^{(2)}E + A_2)(\omega_k^{(3)}E + A_3), \quad \tau_k = \omega_k^{(2)} + \omega_k^{(3)}.$$

Dans ce cas l'opérateur $A_3 = -\Lambda_3$, $A_2y = -\Lambda_2\bar{y}$, où l'opérateur de différences Λ_3 est défini dans (26) avec $d_3 = 0{,}5d$, tandis que $\Lambda_2 y = (a_2 y_{\bar{\varphi}})_{\hat{\varphi}} - 0{,}5dy$. Les constantes δ_α et Δ_α sont les bornes de l'opérateur A_α et s'apprécient de la façon suivante: δ_3 et Δ_3 ont été trouvés au point 3, § 2, la constante δ_2 s'obtient de manière exacte: $\delta_2 = 0{,}5d$, tandis qu'en guise de Δ_2 on peut prendre

$$\Delta_2 = \max_{0 \leqslant j \leqslant N_2 - 1}\left[\frac{2}{\hbar_2(j)}\left(\frac{a_2(j)}{h_2(j)} + \frac{a_2(j+1)}{h_2(j+1)}\right) + \frac{d}{2}\right].$$

§ 3. Résolution des problèmes de différences dans le système de coordonnées polaires

1. Schémas aux différences pour les équations dans un cercle et un anneau. Examinons les méthodes de résolution des schémas aux différences pour des équations elliptiques sans dérivées mixtes dans le système de coordonnées polaires. Etudions d'abord le cas où le domaine dans lequel est recherchée la solution est un cercle ou un anneau dans le système de coordonnées cartésiennes. En coordonnées polaires, aux domaines mentionnés correspond le rectangle $\bar{G} =$

$= \{l_1 \leqslant r \leqslant L_1,\ l_2 \leqslant \varphi \leqslant L_2,\ l_1 \geqslant 0,\ L_2 - l_2 = 2\pi\}$. Il s'agit de trouver la solution de l'équation

$$\frac{1}{r}\frac{\partial}{\partial r}\left(rk_1\frac{\partial u}{\partial r}\right)+\frac{1}{r^2}\frac{\partial}{\partial\varphi}\left(k_2\frac{\partial u}{\partial\varphi}\right)-qu=-f,\quad (r,\ \varphi)\in G, \tag{1}$$

périodique en φ de période 2π et satisfaisant à la frontière du rectangle $\overline{G}$ aux conditions:

1) pour $r = L_1$, $l_2 \leqslant \varphi \leqslant L_2$ soit à des conditions aux limites de première espèce

$$u(r,\ \varphi) = g_1^+(\varphi), \tag{2}$$

soit de deuxième ou de troisième espèce

$$-k_1\frac{\partial u}{\partial r}=\varkappa_1^+u-g_1^+(\varphi); \tag{3}$$

2) pour $r = l_1 > 0$, $l_2 \leqslant \varphi \leqslant L_2$ soit aux conditions aux limites de première espèce

$$u(r,\ \varphi) = g_1^-(\varphi), \tag{4}$$

soit de deuxième ou de troisième espèce

$$k_1\frac{\partial u}{\partial r}=\varkappa_1^-u-g_1^-(\varphi); \tag{5}$$

pour $r=l_1=0$ on pose la condition

$$\lim_{r\to 0} rk_1\frac{\partial u}{\partial r}=0, \tag{6}$$

distinguant la solution limitée.

On suppose que les coefficients satisfont aux conditions $k_1(r,\ \varphi) \geqslant c_1 > 0$, $k_2(r,\ \varphi) \geqslant c_1 > 0$, $q(r,\ \varphi) \geqslant 0$, $\varkappa_1^{\pm}(\varphi) \geqslant 0$.

Etudions des combinaisons quelconques des conditions aux limites (2)-(5). Construisons les schémas aux différences correspondant aux conditions aux limites mentionnées.

Introduisons dans le domaine $\overline{G}$ un maillage irrégulier rectangulaire quelconque

$$\overline{\omega} = \{(r_i,\ \varphi_j) \in \overline{G},\ r_i = r_{i-1} + h_1(i),\ 1 \leqslant i \leqslant N_i,\ r_0 = l_1,$$
$$r_{N_1} = L_1,\ \varphi_j = \varphi_{j-1} + h_2(j),\ 1 \leqslant j < N_2,\ \varphi_0 = l_2,\ \varphi_{N_2} = L_2\}.$$

Le pas moyen $\hbar_1(i)$ est défini au point 1, § 2, tandis que le pas $\hbar_2(j)$ l'est au point 4, § 2, par la formule (45). Définissons la fonction de maille $\rho(i)$:

$$\rho(i)=\begin{cases} l_1+\frac{1}{4}h_1(1), & i=0,\\ r_i+\frac{1}{4}[h_1(i+1)-h_1(i)], & 1\leqslant i\leqslant N_1-1,\\ L_1-\frac{1}{4}h_1(N_1), & i=N_1.\end{cases} \tag{7}$$

Dans le cas primitif des coefficients continus k_1, k_2, q et f les coefficients du schéma aux différences seront déterminés suivant les formules

$$a_1(i,\ j) = \bar{r}_1 k_1(\bar{r}_i,\ \varphi_j),\quad a_2(i,\ j) = k_2(r_i,\ \bar{\varphi}_j),$$

$$d(i,\ j) = q(r_i\varphi_j),\qquad \psi(i,\ j) = f(r_i,\ \varphi_j),$$

où $\bar{r}_i = r_i - 0{,}5h_1(i)$, $\bar{\varphi}_j = \varphi_j - 0{,}5h_2(j)$.

En utilisant les notations introduites, approximons (1) aux équations aux différences

$$\Lambda y = \frac{1}{\rho}(a_1 y_{\bar{r}})_{\hat{r}} + \frac{1}{\rho^2}(a_2 y_{\bar{\varphi}})_{\hat{\varphi}} - dy = -\psi, \tag{8}$$
$$1 \leqslant i \leqslant N_1 - 1,\quad 0 \leqslant j \leqslant N_2 - 1.$$

Afin de rendre plus compacte l'écriture, on a recours aux relations

$$\begin{gathered} y(i,\ j) = y(i,\ N_2 + j),\quad j = 0,\ -1,\quad a_2(i,\ 0) = a_2(i,\ N_2), \\ h_2(0) = h_2(N_2), \end{gathered} \tag{9}$$

qui se dégagent de la condition de périodicité.

Les conditions aux limites (2), (4) s'approximent de façon stricte

$$y(N_1,\ j) = g_1^+(\varphi_j),\qquad y(0,\ j) = g\ \ (\varphi_j),\qquad 0 \leqslant j \leqslant N_2 - 1. \tag{10}$$

L'analogue au sens des différences finies des conditions aux limites de troisième espèce (3), (5) a pour expression (pour $0 \leqslant j \leqslant N_2 - 1$)

$$\Lambda y = -\frac{a_1}{\rho\hbar_1} y_{\bar{r}} + \frac{1}{\rho^2}(a_2 y_{\bar{\varphi}})_{\hat{\varphi}} - \left(d + \frac{r\varkappa_1^+}{\rho\hbar_1}\right) y = \\ = -\psi - \frac{rg_1^+}{\rho\hbar_1},\quad i = N_1, \tag{11}$$

$$\Lambda y = \frac{a_1^{+1}}{\rho\hbar_1} y_r + \frac{1}{\rho^2}(a_2 y_{\bar{\varphi}})_{\hat{\varphi}} - \left(d + \frac{r\varkappa_1^-}{\rho\hbar_1}\right) y = \\ = -\psi - \frac{rq_1^-}{\rho\hbar_1},\quad i = 0. \tag{12}$$

On a utilisé ici les relations (9).

Il reste à construire la condition aux limites au sens des différences finies imposée au côté $r = l_1$ pour le cas où $l_1 = 0$. Comme tous les nœuds portés par le côté $r = 0$ s'identifient, on a

$$y(0,\ j) = y_0,\quad 0 \leqslant j \leqslant N_2 - 1. \tag{13}$$

Vu que l'origine des coordonnées est un point intérieur à un cercle, on obtient alors, en écrivant l'équation (1) dans le système de coordonnées cartésiennes et en l'approximant sur un maillage radialo-

annulaire avec la condition (6),

$$\Lambda y = \frac{1}{2\pi\rho\hbar_1} \sum_{j=0}^{N_2-1} a_1^{+1} y_r \hbar_2 - dy = -\psi, \quad i = 0, \qquad (14)$$

$$d(0,\ j) = d_0, \quad \psi(0,\ j) = \psi_0, \quad 0 \leqslant j \leqslant N_2 - 1.$$

Ici y_0, d_0 et ψ_0 sont les valeurs des fonctions de mailles correspondantes du centre du cercle.

Ainsi, au cas d'un cercle on a une condition aux limites non locale (13), (14) sur le côté $r = 0$ du rectangle $\overline{G}$. Les schémas aux différences sont ainsi construits.

Pour l'approximation au sens des différences finies de l'équation (1) au voisinage de $r = 0$ on utilise souvent un autre maillage en r dans lequel le point $r = 0$ n'est pas compris:

$$\overline{\omega} = \{(r_i,\ \varphi_j) \in \overline{G},\quad r_i = (i + 0{,}5)\, h_1, \quad 0 \leqslant i \leqslant N_1, \quad r_{N_1} = L_1,$$
$$\varphi_j = \varphi_{j-1} + h_2(j), \quad 1 \leqslant j \leqslant N_2, \quad \varphi_0 = l_2, \quad \varphi_{N_2} = L_2\}$$

(pour simplifier, on admet que le maillage est régulier en r).

Dans ce cas $a_1(i,\ j) = \overline{r}_i k_1(\overline{r}_i,\ \varphi_j)$, $a_2(i,\ j) = k_2(r_i,\ \overline{\varphi}_j)$, etc., où $\overline{r}_i = ih_1$. Les équations (8) restent inchangées, et avec $i = 0$ on écrit l'équation aux différences suivante:

$$\Lambda y = \frac{\overline{r}_1}{r_0 h_1} a_1(1,\ j)\, y_r(1,\ j) + \frac{1}{r_0} (a_2 y_{\overline{\varphi}})_{\hat{\varphi}} - dy = -\psi$$

(r_0 est ici égal à $0{,}5h_1$, $\overline{r}_1 = h_1$) qui est un analogue de la condition aux limites de troisième espèce.

La condition disparaît pour $r = 0$; on ne peut donc pas déterminer la valeur de y pour $r = 0$ à partir des équations aux différences.

2. Résolubilité des problèmes aux limites discrets. On a construit au point 1 des schémas aux différences approximant les problèmes (1)-(6). Pour un cercle, le schéma est donné par les formules (8), (10), (11), (14), pour un anneau, par les formules (8), (10), (12). Etudions la question de résolubilité de schémas mentionnés.

Désignons par $\overline{\omega}^*$ une partie du maillage $\overline{\omega}$: $\overline{\omega}^* = \{(r_i,\ \varphi_j) \in \overline{\omega},\ 0 \leqslant i \leqslant N_1,\ 0 \leqslant j \leqslant N_2 - 1\}$. L'espace H est composé des fonctions de mailles associées à $\overline{\omega}^*$ et satisfaisant à la condition supplémentaire $y(0,\ j) = \text{const}$, $0 \leqslant j \leqslant N_2 - 1$, si $l_1 = 0$. Définissons le produit scalaire dans H par la formule

$$(u,\ v) = \sum_{i=0}^{N_1} \sum_{j=0}^{N_2-1} u(i,\ j)\, v(i,\ j)\, \rho(i)\, \hbar_1(i)\, \hbar_2(j).$$

On peut montrer que si la fonction $\rho(i)$ est définie par la formule (7), on a l'égalité

$$(1,\ 1) = 0{,}5\,(L_1^2 - l_1^2)\,(L_2 - l_2) = \pi\,(L_1^2 - l_1^2), \qquad (15)$$

c'est-à-dire que le carré de la norme de la fonction, identiquement égale à l'unité sur $\overline{\omega}^*$, est égal à la surface du cercle ($l_1 = 0$) ou de l'anneau ($l_1 > 0$). En outre, si le domaine envisagé est un cercle, alors, en utilisant la constance en j pour $i = 0$ des fonctions de mailles de H ainsi que l'égalité $\sum_{j=0}^{N_2-1} \hbar_2(j) = L_2 - l_2 = 2\pi$, on peut aboutir à l'expression suivante du produit scalaire introduit plus haut:

$$(u,\ v) = \rho(0)\,\hbar_1(0)\,2\pi u_0 v_0 + \sum_{i=1}^{N_1} \sum_{j=0}^{N_2-1} u(i,\ j)\,v(i,\ j)\,\rho(i)\,\hbar_1(i)\,\hbar_2(j), \quad (16)$$

où $u_0 = u(0,\ j)$, $v_0 = v(0,\ j)$.

Etudions la résolubilité des schémas aux différences (8), (11), (13), (14) pour $l_1 = 0$ et (8), (11), (12) pour $l_1 > 0$, si $d \equiv 0$, $\varkappa_1^+ = \varkappa_1^- \equiv 0$. Ecrivons les problèmes de différences susmentionnés sous forme d'une équation opératorielle

$$Au = f, \quad (17)$$

où l'opérateur A se définit de la façon suivante $Ay = -\Lambda\overline{y}$, $y(i,\ j) = \overline{y}(i,\ j)$ pour $0 \leqslant i \leqslant N_1$, $0 \leqslant j \leqslant N_2 - 1$ et $\overline{y}$ remplit les conditions de périodicité (9), en outre $y(0,\ j) = \overline{y}(0,\ j) = \text{const.}$

Examinons d'abord l'opérateur A correspondant à l'opérateur de différence Λ du problème (8), (11), (13), (14). Compte tenu de ce que la première formule de différences de Green des fonctions satisfaisant à la condition de périodicité (9) prend la forme

$$\sum_{j=0}^{N_2-1} (a_2 u_{\bar{\varphi}})_{\hat{\varphi}}\, v\hbar_2 = -\sum_{j=0}^{N_2-1} a_2 u_{\bar{\varphi}} v_{\bar{\varphi}} h_2,$$

il vient, compte tenu de (16),

$$(Au,\ v) = -(\Lambda\overline{u},\ \overline{v}) =$$

$$= \sum_{j=0}^{N_2-1} \hbar_2 \Big(\sum_{i=1}^{N_2} h_1 a_1 \overline{u}_{\bar{r}} \overline{v}_{\bar{r}} + \sum_{i=0}^{N_1} \rho\hbar_1 d\overline{u}\overline{v} + r\varkappa_1^+ \overline{u}\overline{v}\big|_{i=N_1} \Big) +$$

$$+ \sum_{i=1}^{N_1} \frac{\hbar_1}{\rho} \sum_{j=0}^{N_2-1} h_2 a_2 \overline{u}_{\bar{\varphi}} \overline{v}_{\bar{\varphi}} = -(\overline{u},\ \Lambda\overline{v}) = (u,\ Av).$$

Par conséquent, l'opérateur A est autoadjoint dans H.

Pour l'opérateur A correspondant à l'opérateur de différence Λ du problème (8), (11), (12) on obtient une égalité analogue

$$(Au, v) = \sum_{j=0}^{N_2-1} \hbar_2 \left(\sum_{i=1}^{N_1} h_1 a_1 \bar{u}_{\bar{r}} \bar{v}_{\bar{r}} + \sum_{i=0}^{N_1} \rho \hbar_1 d\bar{u}\bar{v} + r\varkappa_1^- \bar{u}\bar{v}|_{i=0} + \right.$$
$$\left. + r\varkappa_1^+ \bar{u}\bar{v}|_{i=N_1} \right) + \sum_{i=0}^{N_1} \frac{\hbar_1}{\rho} \sum_{j=0}^{N_2-1} h_2 a_2 \bar{u}_{\bar{\varphi}} \bar{v}_{\bar{\varphi}} = (u, Av),$$

à partir de laquelle il s'ensuit que l'opérateur A est autoadjoint.

Si $d \equiv 0$, $\varkappa_1^{\pm} \equiv 0$, il s'ensuit de ce que l'opérateur A est autoadjoint et de l'inégalité de Cauchy-Bouniakovski $(Au, u) \leqslant \leqslant \| Au \| \| u \|$ que le noyau de l'opérateur A est composé de fonctions de mailles égales aux constantes du maillage $\bar{\omega}^*$. Aussi la condition de l'existence de la solution de l'équation (17) prend-elle la forme $(f, 1) = 0$. Pour le problème (8), (11), (13), (14) il lui correspond la condition

$$\sum_{i=0}^{N_1} \sum_{j=0}^{N_2-1} \psi(i, j) \rho(i) \hbar_1(i) \hbar_2(j) + L_1 \sum_{j=0}^{N_2-1} \hbar_2(j) g_1^+(\varphi_j) = 0 \quad (18)$$

qui est l'analogue au sens des différences finies de la condition (23) du § 1. Pour le problème (8), (11), (12) la condition de résolubilité est de la forme

$$\sum_{i=0}^{N_1} \sum_{j=0}^{N_2-1} \psi(i, j) \rho(i) \hbar_1(i) \hbar_2(j) + \sum_{j=0}^{N_2-1} \hbar_2(j) \left[L_1 g_1^+(\varphi_j) + l_1 g_1^-(\varphi_j) \right] = 0$$

et est un analogue de la condition (25) du § 1, qui garantit la résolubilité du problème différentiel correspondant pour l'anneau.

Si les conditions mentionnées sont remplies, les solutions des problèmes examinés existent bien et deux quelconques de ces solutions diffèrent d'une constante. La solution normale de ces problèmes satisfait à la condition $(\bar{y}, 1) = 0$.

Soit y l'une de ces solutions qu'on arrive à trouver, par exemple, en fixant la solution cherchée en un point du maillage. Alors en tenant compte de l'égalité (15) on obtient la fonction

$$\bar{y} = y - \frac{(y, 1)}{\pi (L_1^2 - l_1^2)} = y - \frac{(y, 1)}{(1, 1)}$$

qui est une solution normale.

R e m a r q u e. Si l'on définit la fonction de maille $\rho(i)$ au moyen des formules

$$\rho(i) = r_i, \ 1 \leqslant i \leqslant N_1, \ \rho(0) = \begin{cases} h_1(0)/4, & l_1 = 0, \\ l_1, & l_1 > 0, \end{cases}$$

seule l'égalité (15) changera au cas où $l_1 = 0$. On aura alors

$$(1,\ 1)=\pi L_1^2+\frac{h_1^2(1)}{4}\pi=\pi\left(L_1^2+\frac{h_1^2(1)}{4}\right).$$

3. Principe de superposition pour le problème dans un cercle. La résolution des problèmes de différences dans un cercle se heurte à l'existence de la condition aux limites non locale (14) qui survient lorsque $i=0$. Notons que si le problème est dégénéré et la condition de résolubilité (18) est remplie, il est alors commode de séparer une des solutions en fixant sa valeur au centre du cercle, c'est-à-dire en posant $y(0,\ j) = y_0$; $0 \leqslant j \leqslant N_2 - 1$. Dans ce cas la condition (14) est écartée et le problème obtenu avec y_0 donné est analogue à celui posé pour l'anneau avec des conditions aux limites de première espèce sur le cercle intérieur. Supposons maintenant que le problème de différences (8), (11), (13), (14) n'est pas dégénéré. Montrons qu'il est possible d'obtenir sa solution en résolvant deux problèmes auxiliaires avec conditions aux limites locales de première espèce pour $i = 0,\ 0 \leqslant j \leqslant N_2 - 1$.

Cherchons la solution du problème (8), (11), (13), (14) sous la forme

$$y(i,\ j) = v(i,\ j) + y_0 w(i,\ j), \quad 0 \leqslant i \leqslant N_1, \quad 0 \leqslant j \leqslant N_2 - 1, \tag{19}$$

où y_0 est la valeur de la solution cherchée au centre du cercle, tandis que $v(i,\ j)$ et $w(i,\ j)$ satisfont aux conditions de périodicité

$$v(i,\ j) = v(i,\ N_2 + j), \quad w(i,\ j) = w(i,\ N_2 + j), \quad j = 0,\ -1$$

et sont des solutions des problèmes aux limites suivants:

$$\left.\begin{aligned}&\frac{1}{\rho}(a_1 v_{\bar r})_{\hat r}+\frac{1}{\rho^2}(a_2 v_{\bar\varphi})_{\hat\varphi}-dv=-\psi, \quad 1\leqslant i\leqslant N_1-1,\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad 0\leqslant j\leqslant N_2-1,\\ &v(0,\ j)=0, \quad i=0,\end{aligned}\right\} \tag{20}$$

$$-\frac{a_1}{\rho\hbar_1}v_{\bar r}+\frac{1}{\rho^2}(a_2 v_{\bar\varphi})_{\hat\varphi}-\left(d+\frac{r\varkappa_1^+}{\rho\hbar_1}\right)v=-\psi-\frac{rg_1^+}{\rho\hbar_1}, \quad i=N_1,$$

$$\left.\begin{aligned}&\Lambda w=\frac{1}{\rho}(a_1 w_{\bar r})_{\hat r}+\frac{1}{\rho^2}(a_2 w_{\bar\varphi})_{\hat\varphi}-dw=0, \quad 1\leqslant i\leqslant N_1-1,\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad 0\leqslant j\leqslant N_2-1,\\ &w(0,\ j)=1, \quad i=0,\\ &-\frac{a_1}{\rho\hbar_1}w_{\bar r}+\frac{1}{\rho^2}(a_2 w_{\bar\varphi})_{\hat\varphi}-\left(d+\frac{r\varkappa_1^+}{\rho\hbar_1}\right)w=0, \quad i=N_1.\end{aligned}\right\} \tag{21}$$

La fonction y définie suivant (19) satisfait apparemment à l'équation (8) et aux conditions (11), (13). Il reste à déterminer y_0. En

portant (19) dans la condition (14) qui n'était pas encore utilisée, et en tenant compte des conditions aux limites pour v et w, il vient

$$y_0 = \frac{\left[2\pi\rho\hbar_1\psi_0 + \sum_{j=0}^{N_2-1} a_1^{+1} v_r \hbar_2(j)\right]_{i=0}}{\left[2\pi\rho\hbar_1 d_0 - \sum_{j=0}^{N_2-1} a_1^{+1} w_r \hbar_2(j)\right]_{i=0}}. \qquad (22)$$

Montrons que le dénominateur dans (22) est différent de zéro. A cette fin multiplions l'équation (21) scalairement par w. En recourant aux conditions aux limites pour w, aux relations de périodicité et aux formules de différences de Green, on obtient

$$0 = \sum_{i=1}^{N_1-1} \sum_{j=0}^{N_2-1} (\Lambda w)\, w\rho\hbar_1\hbar_2 = -\sum_{j=0}^{N_2-1} \hbar_2 \left(a_1^{+1} w_r)_{i=0} + L_1 \varkappa_1^+ w^2|_{i=N_1}\right) -$$

$$-\sum_{i=1}^{N_1} \sum_{j=0}^{N_2-1} a_1 w_{\bar{r}}^2 h_1 \hbar_2 - \sum_{i=1}^{N_1} \sum_{j=0}^{N_2-1} \hbar_1 \left[\frac{h_2}{\rho} a_2 w_{\bar{\varphi}}^2 + \hbar_2 \rho\, d w^2\right].$$

Vu que la fonction w n'est pas une constante, $d \geqslant 0$, $a_\alpha \geqslant c_1 > 0$, $\alpha = 1, 2$, et $\varkappa_1^+ \geqslant 0$, avec $d^2 + (\varkappa_1^+)^2 \neq 0$, il s'ensuit que

$$\sum_{j=0}^{N_2-1} a_1^{+1} w_r \hbar_2|_{i=0} < 0$$

et, par conséquent, le dénominateur dans la formule (22) est différent de zéro.

La résolution du problème de départ (8), (11), (13), (14) est donc réduite à la résolution de deux problèmes (20) et (21) avec conditions aux limites locales et à l'obtention de y_0 suivant la formule (22). La solution cherchée de y s'obtient à l'aide de la formule (19).

Notons que si au côté $r = L_1$ est imposée la condition aux limites de première espèce $y(N_1, j) = g_1^+(\varphi_j)$, alors pour les fonctions v et w, au lieu des conditions de troisième espèce, il faut exiger dans (20) et (21) $v(N_1, j) = g_1^+(\varphi_j)$ et $w(N_1, j) = 0$ pour $0 \leqslant j \leqslant N_2 - 1$. La formule (22) pour y_0 se conserve. Si les coefficients k_1, k_2, q et $\varkappa_1^+$ sont indépendants de φ, la solution w du problème (21) est alors également indépendante de φ. Dans ce cas pour la fonction w on est en présence d'un problème unidimensionnel

$$\frac{1}{\rho}(a_1 w_{\bar{r}})_{\hat{r}} - dw = 0, \quad 1 \leqslant i \leqslant N_1 - 1,$$

$$w(0, j) = 1, \ i = 0,$$

$$-\frac{a_1}{\rho\hbar_1} w_{\bar{r}} - \left(d + \frac{r\varkappa_1^+}{\rho\hbar_1}\right) w = 0, \quad i = N_1,$$

qui se résout par la méthode du balayage.

4. Méthodes directes de résolution des équations dans le cercle et l'anneau. Il s'ensuit de ce qui a été dit plus haut qu il suffit de se borner à l'étude des méthodes de résolution des problèmes de différences (8), (10)-(12). Etudions d'abord le cas pour lequel les problèmes de différences mentionnés se prêtent à la résolution par l'une des méthodes directes exposées dans les chapitres III et IV.

Supposons que les coefficients k_1, k_2 et q de l'équation (1) ne dépendent pas de φ: $k_1 = k_1(r)$, $k_2 = k_2(r)$, $q = q(r)$. C'est la situation qui se présente pour l'équation de Poisson en coordonnées polaires. De plus, admettons que dans les conditions aux limites de troisième espèce (11), (12) $\varkappa_1^-$ et $\varkappa_1^+$ sont des constantes. On suppose que le maillage $\bar{\omega}$ est régulier en φ, c'est-à-dire que $h_2(j) \equiv h_2$, et peut de même être irrégulier en r. Sous les hypothèses admises, l'équation aux différences (8) avec une combinaison quelconque des conditions aux limites (10)-(12) se prête à la résolution soit par la méthode de réduction totale, soit par la méthode combinée de réduction incomplète et de séparation des variables.

Illustrons la possibilité d'application des méthodes directes par un exemple dans lequel aux côtés $r = l_1$ et $r = L_1$ sont imposées des conditions aux limites de troisième (de deuxième) espèce (11), (12). Les autres combinaisons des conditions aux limites sont étudiées de façon analogue.

En vertu des hypothèses émises les coefficients du schéma aux différences se déterminent par les formules

$$a_1(i) = \bar{r}_i k_1(\bar{r}_i), \quad a_2(i) = k_2(r_i), \quad d(i) = q(r_i),$$

et comme le maillage $\bar{\omega}$ est régulier en φ, l'opérateur de différences $(a_2 y_{\bar{\varphi}})_{\hat{\varphi}}$ est remplacé par $a_2 y_{\bar{\varphi}\varphi}$.

Ramenons le problème de différences (8), (11), (12) à un système d'équations vectorielles triponctuelles

$$\begin{aligned} -Y_{N_2-1} + CY_0 - Y_1 &= F_0, & j &= 0, \\ -Y_{j-1} + CY_j - Y_{j+1} &= F_j, & 1 &\leqslant j \leqslant N_2 - 2, \\ -Y_{N_2-2} + CY_{N_2-1} - Y_0 &= F_{N_2-1}, & j &= N_2 - 1. \end{aligned} \qquad (23)$$

On a utilisé ici pour $0 \leqslant j \leqslant N_2 - 1$ les notations:

$$Y_j = (y(0, j), y(1, j), \ldots, y(N_1, j)),$$

$$F_j = (\theta_0 f(0, j), \theta_1 f(1, j), \ldots, \theta_{N_1} f(N_1, j)),$$

$$CY_j = ((2E - \theta_0 \Lambda_1) y(0, j), \ldots, (2E - \theta_{N_1} \Lambda_1) y(N_1, j)),$$

où

$$f(i,\ j)=\begin{cases}\psi(0,\ j)+\dfrac{l_1 g_1^-(\varphi_j)}{\rho(0)\,\hbar_1(0)}, & i=0,\\ \psi(i,\ j), & 1\leqslant i\leqslant N_1-1,\\ \psi(N_1,\ j)+\dfrac{L_1 g_1^+(\varphi_j)}{\rho(N_1)\,\hbar_1(N_1)}, & i=N_1,\end{cases}\tag{24}$$

est l'opérateur de différences Λ_1 qui agit de la façon suivante:

$$\Lambda_1 y=\begin{cases}\dfrac{a_1^{+1}}{\rho\hbar_1}y_r-\left(d+\dfrac{r\varkappa_1^-}{\rho\hbar_1}\right)y, & i=0,\\ \dfrac{1}{\rho}(a_1 y_{\bar r})_{\hat r}-dy, & 1\leqslant i\leqslant N_1-1,\\ -\dfrac{a_1}{\rho\hbar_1}y_{\bar r}-\left(d+\dfrac{r\varkappa_1^+}{\rho\hbar_1}\right)y, & i=N_1,\end{cases}\tag{25}$$

et, enfin, $\theta_i = \rho^2(i)\,h_2^2/a_2(i)$, $0 \leqslant i \leqslant N_1$.

Le système (23) s'obtient à partir de (8), (11) et (12) au moyen de la multiplication de chaque équation par θ_i correspondant et passage à l'écriture vectorielle.

Rappelons que l'algorithme de la méthode de réduction totale est décrit pour le système (23) au point 2, § 4, ch. III. Dans la méthode combinée on utilise l'algorithme de la transformation discrète rapide de Fourier donné au point 4, § 1, ch. IV. Ces méthodes se caractérisent par l'estimation des opérations arithmétiques se montant à $O(N_1N_2\log_2 N_2)$ avec $N_2 = 2^n$.

5. Méthode des directions alternées. Supposons maintenant que les coefficients de l'équation (1) et des conditions aux limites (3), (5) satisfont aux conditions $k_1 = k_1(r)$, $k_2 = k_2(\varphi)$, $q = \text{const}$, $\varkappa_1^{\pm} = \text{const}$, c'est-à-dire que pour le problème (1), (3), (5) la méthode de séparation des variables est applicable. On admet que le maillage $\bar\omega$ est irrégulier dans chaque direction. Etudions l'équation aux différences (8) avec une combinaison quelconque des conditions aux limites (10)-(12). Sous les hypothèses émises, les variables du schéma aux différences se séparent et sa solution approchée peut être obtenue avec la méthode des directions alternées munie d'un jeu optimal de paramètres d'itération.

En guise d'exemple, examinons le problème (8), (11), (12) avec les conditions aux limites de troisième espèce pour $r = l_1$ et $r = L_1$. Ecrivons ce problème sous la forme

$$\begin{gathered}\bar\Lambda y = -\bar f,\ 0 \leqslant i \leqslant N_1,\ 0 \leqslant j \leqslant N_2-1,\\ \bar\Lambda = \bar\Lambda_1 + \bar\Lambda_2,\ \bar f = \rho^2 f,\end{gathered}\tag{26}$$

où $\bar{\Lambda}_1 = \rho^2\Lambda_1$, l'opérateur Λ_1 est défini dans (25), l'opérateur $\bar{\Lambda}_2$ donné par l'égalité $\bar{\Lambda}_2 y = (a_2 y_{\bar{\varphi}})_{\hat{\varphi}}$ et les relations (9) satisfaites, tandis que le second membre f est défini dans (24). L'équation (26) est obtenue à partir de (8), (11), (12) après multiplication par ρ^2.

En vertu des hypothèses émises les coefficients du schéma aux différences (26) sont choisis suivant les formules $a_1(i) = \bar{r}_i k_1(\bar{r}_i)$, $a_2(j) = k_2(\bar{\varphi}_j)$, $d = q =$ const.

Dans l'espace H des fonctions de mailles associées à $\bar{\omega}^*$ définissons le produit scalaire

$$(u, v) = \sum_{i=0}^{N_1} \sum_{j=0}^{N_2-1} \frac{\hbar_1(i)\hbar_2(j)}{\rho(i)} u(i, j) v(i, j). \tag{27}$$

Les opérateurs A_1 et A_2 agissant dans H seront définis comme d'habitude: $A_\alpha y = -\Lambda_\alpha \bar{y}$, où $y(i, j) = \bar{y}(i, j)$ pour $0 \leqslant i \leqslant N_1$, $0 \leqslant j \leqslant N_2 - 1$ et $\bar{y}$ remplit les relations de périodicité (9). Dans ce cas le schéma (26) peut s'écrire sous forme d'une équation opératorielle

$$Au = \bar{f}, \quad A = A_1 + A_2 \tag{28}$$

dans l'espace H.

Pour la résolution de l'équation (28) utilisons la méthode des directions alternées, dont le schéma itératif est de la forme

$$\begin{gathered} B_{k+1}\frac{y_{k+1}-y_k}{\tau_{k+1}} + Ay_k = \bar{f}, \quad k = 0, 1, \ldots, y_0 \in H, \\ B_k = (\omega_k^{(1)}E + A_1)(\omega_k^{(2)}E + A_2), \quad \tau_k = \omega_k^{(1)} + \omega_k^{(2)}. \end{gathered} \tag{29}$$

Le fait que les opérateurs A_1 et A_2 sont autoadjoints dans l'espace H s'établit au moyen de la formule de différences de Green, tandis que leur permutabilité se vérifie directement.

Cherchons à présent les bornes des opérateurs A_1 et A_2, c'est-à-dire les constantes δ_α et Δ_α, $\alpha = 1, 2$, des inégalités

$$\delta_\alpha (u, u) \leqslant (A_\alpha u, u) \leqslant \Delta_\alpha (u, u).$$

Cherchons d'abord δ_2 et Δ_2. Vu que pour la fonction $\bar{u}(i, j)$ satisfaisant à la condition de périodicité (9) on a

$$(A_2 u, u) = -(\Lambda_2 \bar{u}, \bar{u}) = \sum_{i=0}^{N_1} \sum_{j=0}^{N_2-1} \frac{\hbar_1(i) h_2(j)}{\rho(i)} (a_2 \bar{u}_{\bar{\varphi}}^2)_{ij},$$

il s'ensuit que

$$\delta_2 = 0, \quad \Delta_2 = \max_{0 \leqslant j \leqslant N_2-1} \left[\frac{a_2(j+1)}{h_2(j+1)} + \frac{a_2(j)}{h_2(j)} \right] \frac{2}{\hbar_2(j)}.$$

On a tenu compte ici des relations (9) pour a_2 et h_2.

Ensuite, en utilisant l'analogue du lemme 16, ch. V, on trouve que δ_1 peut être apprécié de la façon suivante: $1/\delta_1 = \max\limits_{0\leqslant i\leqslant N_1} v(i)$, où $v(i)$ est la solution du problème aux limites

$$\begin{gathered}\rho(a_1 v_{\bar{r}})_{\hat{r}} - d\rho^2 v = -1, \quad 1\leqslant i\leqslant N_1-1,\\ \frac{\rho a_1^{+1}}{\hbar_1} v_r - \left(d+\frac{r\varkappa_1^-}{\rho\hbar_1}\right)\rho^2 v = -1, \quad i=0,\\ -\frac{\rho a_1}{\hbar_1} v_{\bar{r}} - \left(d+\frac{r\varkappa_1^+}{\rho\hbar_1}\right)\rho^2 v = -1, \quad i=N_1.\end{gathered}\tag{30}$$

Le problème (30) se résout par la méthode du balayage.

Obtenons maintenant l'estimation pour Δ_1. En utilisant la première formule de différences de Green et la définition (27) du produit scalaire, il vient

$$(A_1 u, u) = -(\bar{\Lambda}_1\bar{u}, \bar{u}) =$$
$$= \sum_{j=0}^{N_2-1} \hbar_2(j)\Big[\sum_{i=1}^{N_1} h_1(i)\, a_1(i)\, \bar{u}_{\bar{r}}^2(i, j) + d\sum_{i=0}^{N_1} \hbar_1(i)\,\rho(i)\,\bar{u}^2(i, j) +$$
$$+ l_1\varkappa_1^-\bar{u}^2(0, j) + L_1\varkappa_1^+\bar{u}^2(N_1, j)\Big].$$

Apprécions l'expression entre crochets. A partir de l'analogue du lemme 16, ch. V, on obtient l'estimation

$$d\sum_{i=0}^{N_1} \hbar_1(i)\,\rho(i)\,\bar{u}^2(i, j) + l_1\varkappa_1^-\bar{u}^2(0, j) + L_1\varkappa_1^+\bar{u}^2(N_1, j) \leqslant$$
$$\leqslant m_1\Big[\sum_{i=1}^{N_1} (a_1\bar{u}_{\bar{r}}^2)_{ij}\, h_1(i) + \sum_{i=0}^{N_1} \frac{\hbar_1(i)}{\rho(i)}\,\bar{u}^2(i, j)\Big], \tag{31}$$

où $m_1 = \max\limits_{0\leqslant i\leqslant N_1} w(i)$, tandis que $w(i)$ est la solution du problème

$$\begin{gathered}\rho(a_1 w_{\bar{r}})_{\hat{r}} - w = -d\rho^2, \quad 1\leqslant i\leqslant N_1-1,\\ \frac{\rho a_1^{+1}}{\hbar_1} w_r - w = -\left(d+\frac{r\varkappa_1^-}{\rho\hbar_1}\right)\rho^2, \quad i=0,\\ -\frac{\rho a_1}{\hbar_1} w_{\bar{r}} - w = -\left(d+\frac{r\varkappa_1^+}{\rho\hbar_1}\right)\rho^2, \quad i=N_1.\end{gathered}\tag{32}$$

Ensuite, de l'analogue du lemme 17, ch. V, on obtient l'estimation

$$\sum_{i=1}^{N_1} a_1(i)\,\bar{u}_{\bar{r}}^2(i, j)\, h_1(i) \leqslant m_2 \sum_{i=0}^{N_1} \frac{\hbar_1(i)}{\rho(i)}\,\bar{u}^2(i, j), \tag{33}$$

où

$$m_2 = \max\left(\frac{a_1(N_1)\rho(N_1)}{\hbar_1^2(N_1)}, \frac{a_1(1)\rho(0)}{\hbar_1^2(0)}, \max_{1\leqslant i\leqslant N_1-1} \frac{2\rho(i)}{\hbar_1(i)}\left[\frac{a_1(i)}{h_1(i)} + \frac{a_1(i+1)}{h_1(i+1)}\right]\right).$$

De (31) et (33) on déduit l'estimation

$$\sum_{i=1}^{N_1} h_1 a_1 \bar{u}_{\bar{r}}^2 + d \sum_{i=0}^{N} \hbar_1 \rho \bar{u}^2 + l_1 \varkappa_1^- \bar{u}^2 \,|_{i=0} + L_1 \varkappa_1^+ \bar{u}^2 \,|_{i=N_1} \leqslant$$

$$\leqslant \Delta_1 \sum_{i=0}^{N_1} \frac{\hbar_1}{\rho} \bar{u}^2, \qquad \Delta_1 = m_1 + m_2 (1 + m_1).$$

En multipliant cette inégalité par $\hbar_2(j)$ et en sommant en j de 0 à $N_2 - 1$, on obtient $(A_1 u, u) \leqslant \Delta_1 (u, u)$.

Ainsi, les constantes δ_α et Δ_α, $\alpha = 1, 2$, sont obtenues. Rappelons que les formules des paramètres d'itération $\omega_k^{(1)}$ et $\omega_k^{(2)}$ ont été obtenues au point 4, § 1, ch. XI.

On construit de façon analogue la méthode des directions alternées pour le problème de différences (8), (10) avec conditions aux limites de première espèce. Les constantes δ_α et Δ_α s'apprécient de la même façon qu'au cas étudié auparavant, il faut seulement remplacer dans (30) et (32) les conditions aux limites de troisième espèce par les conditions $v(0) = 0$, $v(N_1) = 0$ et $w(0) = 0$, $w(N_1) = 0$.

Notons en conclusion que pour $d = 0$, $\varkappa_1^\pm = 0$ le problème (8), (11), (12) est dégénéré, et si la condition de résolubilité

$$\sum_{i=0}^{N_1} \sum_{j=0}^{N_2-1} \psi \rho \hbar_1 \hbar_2 + \sum_{j=0}^{N_2-1} \hbar_2 [L_1 g_1^+ + l_1 g_1^-] = 0$$

est remplie le problème présente une solution non unique. Dans ce cas le jeu de paramètres $\omega_k^{(1)}$ et $\omega_k^{(2)}$ pour la méthode des directions alternées (29) a été construit au point 1, § 4, ch. XII.

6. Résolution des problèmes de différences dans un secteur annulaire. Examinons les méthodes de résolution des problèmes de différences aux limites pour l'équation elliptique sans dérivées mixtes et donnée dans un secteur annulaire.

Dans le domaine $\bar{G} = \{l_1 \leqslant r \leqslant L_1,\ l_2 \leqslant \varphi \leqslant L_2,\ l_1 > 0,\ L_2 - l_2 < 2\pi\}$ il s'agit de trouver la solution des équations (1) qui satisfait sur les côtés $r = l_1$ et $r = L_1$ à l'une des conditions aux limites (2)-(5) et sur les côtés $\varphi = l_2$ et $\varphi = L_2$ à l'une des conditions

$$u(r, \varphi) = g_2^-(r), \quad \varphi = l_2 \tag{34}$$

ou bien

$$\frac{k_2}{r} \frac{\partial u}{\partial \varphi} = \varkappa_2^- u - g_2^+(r), \quad \varphi = l_2, \tag{35}$$

$$u(r, \varphi) = g_2^+(r), \qquad \varphi = L_2 \tag{36}$$

ou bien

$$-\frac{k_2}{r} \frac{\partial u}{\partial \varphi} = \varkappa_2^+ u - g_2^+(r), \quad \varphi = L_2. \tag{37}$$

On admet que les coefficients satisfont aux conditions $k_1(r, \varphi) \geqslant \geqslant c_1 > 0$, $k_2(r, \varphi) \geqslant c_1 > 0$, $q(r, \varphi) \geqslant 0$, $\varkappa_1^\pm(\varphi) \geqslant 0$, $\varkappa_2^\pm(r) \geqslant 0$.

On introduit dans le domaine $\bar{G}$ un maillage rectangulaire irrégulier quelconque $\bar{\omega}$ (voir point 1, § 3):

$$\bar{\omega} = \{(r_i, \varphi_j) \in \bar{G},\ r_i = r_{i-1} + h_1(i),\ 1 \leqslant i \leqslant N_1,\quad r_0 = l_1,\ r_{N_1} = L_1,\ \varphi_j = \varphi_{j-1} + h_2(j),\ 1 \leqslant j \leqslant N_2,\ \varphi_0 = l_2,\ \varphi_{N_2} = L_2\}$$

et l'on détermine les pas moyens $\hbar_1(i)$ et $\hbar_2(j)$:

$$\hbar_\alpha(m) = \begin{cases} 0{,}5h_\alpha(1), & m = 0, \\ 0{,}5[h_\alpha(m) + h_\alpha(m+1)], & 1 \leqslant m \leqslant N_\alpha - 1, \\ 0{,}5h_\alpha(N_\alpha), & m = N_\alpha, \quad \alpha = 1, 2. \end{cases}$$

L'équation (1) est approximée par l'équation aux différences

$$\frac{1}{\rho}(a_1 y_{\bar{r}})_{\hat{r}} + \frac{1}{\rho^2}(a_2 y_{\bar{\varphi}})_{\hat{\varphi}} - dy = -\psi, \quad (38)$$
$$1 \leqslant i \leqslant N_1 - 1, \quad 1 \leqslant j \leqslant N_2 - 1.$$

Les conditions aux limites de première espèce (2), (4), (34), (36) sont approximées de façon stricte:

$$y(N_1, j) = g_1^+(\varphi_j), \quad y(0, j) = g_1^-(\varphi_j), \ 0 \leqslant j \leqslant N_2, \quad (39)$$
$$y(i, N_2) = g_2^+(r_i), \quad y(i, 0) = g_2^-(r_i), \ 0 \leqslant i \leqslant N_1. \quad (40)$$

Les conditions de troisième espèce (3) et (5), données avec $r = L_1$ et $r = l_1$, sont remplacées pour $1 \leqslant j \leqslant N_2 - 1$ par les conditions (11) et (12).

L'analogue au sens des différences finies des conditions aux limites (35) et (37) est de la forme

$$\frac{1}{\rho}(a_1 y_{\bar{r}})_{\hat{r}} + \frac{a_2^{+1}}{\rho^2 \hbar_2} y_\varphi - \left(d + \frac{\varkappa_2^-}{\rho \hbar_2}\right) y = -\psi - \frac{g_2^-}{\rho \hbar_2}, \quad j = 0, \quad (41)$$
$$\frac{1}{\rho}(a_1 y_{\bar{r}})_{\hat{r}} - \frac{a_2}{\rho^2 \hbar_2} y_{\bar{\varphi}} - \left(d + \frac{\varkappa_2^+}{\rho \hbar_2}\right) y = -\psi - \frac{g_2^+}{\rho \hbar_2}, \quad j = N_2. \quad (42)$$

Si aux côtés qui se coupent du rectangle on impose les conditions aux limites de troisième espèce, il faut poser pour les nœuds d'angles du maillage $\bar{\omega}$ les conditions aux limites suivantes:

$$\frac{a_1^{+1}}{\rho \hbar_1} y_r + \frac{a_2^{+1}}{\rho^2 \hbar_2} y_\varphi - \left(d + \frac{r\varkappa_1^-}{\rho \hbar_1} + \frac{\varkappa_2^-}{\rho \hbar_2}\right) y = -\psi - \frac{r g_1^-}{\rho \hbar_1} - \frac{g_2^-}{\rho \hbar_2}, \quad (43)$$

si $i = j = 0$;

$$-\frac{a_1}{\rho \hbar_1} y_{\bar{r}} + \frac{a_2^{+1}}{\rho^2 \hbar_2} y_\varphi - \left(d + \frac{r\varkappa_1^+}{\rho \hbar_1} + \frac{\varkappa_2^-}{\rho \hbar_2}\right) y = -\psi - \frac{r g_1^+}{\rho \hbar_1} - \frac{g_2^-}{\rho \hbar_2}, \quad (44)$$

si $i = N_1$, $j = 0$:

$$\frac{a_1^{+1}}{\rho \hbar_1} y_r - \frac{a_2}{\rho^2 \hbar_2} y_{\bar{\varphi}} - \left(d + \frac{r\varkappa_1^-}{\rho \hbar_1} + \frac{\varkappa_2^+}{\rho \hbar_2}\right) y = -\psi - \frac{r g_1^-}{\rho \hbar_1} - \frac{g_2^+}{\rho \hbar_2}, \quad (45)$$

si $i=0$, $j=N_2$; et enfin,

$$-\frac{a_1}{\rho\hbar_1} y_{\bar{r}} - \frac{a_2}{\rho^2\hbar_2} y_{\bar{\varphi}} - \left(d + \frac{r\varkappa_1^+}{\rho\hbar_1} + \frac{\varkappa_2^+}{\rho\hbar_2}\right) y =$$
$$= -\psi - \frac{rg_1^+}{\rho\hbar_1} - \frac{g_2^+}{\rho\hbar_2}, \quad (46)$$

si $i = N_1$, $j = N_2$.

Si l'on a le problème de différences (38), (11), (12), (41)-(46) avec $d \equiv 0$ et $\varkappa_\alpha^\pm \equiv 0$, $\alpha = 1, 2$, la solution existe sous réserve de la condition

$$\sum_{j=0}^{N_2} \sum_{i=0}^{N_1} \rho\hbar_1\hbar_2\psi + \sum_{j=0}^{N_2} \hbar_2 (L_1 g_1^+ + l_1 g_1^-) + \sum_{i=0}^{N_1} \hbar_1 (g_2^- + g_2^+) = 0,$$

qui est l'analogue au sens des différences finies de la condition (27) du § 1 de résolubilité du problème respectif pour l'équation différentielle. De plus, deux solutions quelconques du problème mentionné diffèrent d'une constante.

La proposition avancée se démontre presque de la même manière qu'au point 2, § 3 pour le cas du cercle et de l'anneau. Dans l'espace H des fonctions de mailles associées à $\bar{\omega}$ le produit scalaire se définit par la formule

$$(u, v) = \sum_{i=0}^{N_1} \sum_{j=0}^{N_2} u(i, j)\, v(i, j)\, \rho(i)\, \hbar_1(i)\, \hbar_2(j). \quad (47)$$

Notons que les coefficients a_1, a_2, q et la fonction $\rho(i)$ se déterminent en ce point comme au point 1 du § 3.

Faisons une remarque relativement aux méthodes de résolution des problèmes de différences construits. Si les coefficients k_1, k_2, q ne dépendent que de r, $\varkappa_1^\pm$ sont des constantes et $\varkappa_2^\pm = 0$, au cas où sont données les conditions aux limites (3), (5), (35), (37), et que le maillage $\bar{\omega}$ soit régulier en φ, alors les problèmes de différences correspondants peuvent être résolus par des méthodes directes construites dans les chapitres III et IV.

Si sont remplies les conditions $k_1 = k_1(r)$, $k_2 = k_2(\varphi)$, $q = \text{const}$, $\varkappa_2^\pm = \text{const}$ et le maillage $\bar{\omega}$ est irrégulier suivant chaque direction, on peut alors utiliser pour la résolution des problèmes de différences la méthode des directions alternées avec un jeu optimal de paramètres. Dans ce cas, comme il a été fait au point précédent, il est nécessaire de multiplier au préalable les équations aux différences par $\rho^2(i)$.

7. Cas général des coefficients variables. Examinons maintenant le cas où les variables ne se séparent pas et la solution du problème aux limites discret s'obtient par la méthode itérative.

Supposons, par exemple, qu'il s'agit de trouver la solution du problème de Dirichlet pour l'équation (1) sur le maillage $\bar{\omega}$ avec les hypothèses que le maillage $\bar{\omega}$ est régulier en φ ($h_2(j) \equiv h_2$), $q = 0$, tandis que les coefficients k_1 et k_2 satisfont aux conditions

$$0 < c_1 \leqslant k_\alpha(r, \varphi) \leqslant c_2, \quad \alpha = 1, 2. \tag{48}$$

Avec ces hypothèses le problème de différences s'écrit sous la forme

$$\begin{gathered} \Lambda y = \frac{1}{\rho}(a_1 y_{\bar{r}})_{\hat{r}} + \frac{1}{\rho^2}(a_2 y_{\bar{\varphi}})_{\varphi} = -\psi, \quad (r, \varphi) \in \omega, \\ y(r, \varphi) = g(r, \varphi), \quad (r, \varphi) \in \gamma, \end{gathered} \tag{49}$$

où

$$\begin{gathered} a_1(i, j) = \bar{r}_i k_1(\bar{r}_i, \varphi_j), \quad a_2(i, j) = k_2(r_i, \bar{\varphi}_j), \\ \bar{r}_i = r_i - 0{,}5 h_1(i), \quad \bar{\varphi}_j = \varphi_j - 0{,}5 h_2. \end{gathered} \tag{50}$$

Dans l'espace H des fonctions de mailles associées à ω définissons le produit scalaire

$$(u, v) = \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} u(i, j)\, v(i, j)\, \rho(i)\, \hbar_1(i)\, h_2,$$

ainsi que les opérateurs A et R agissant dans H, $Ay = -\Lambda \overset{\bullet}{y}$, $Ry = -\mathscr{R}\overset{\circ}{y}$, où $y(r, \varphi) = \overset{\bullet}{y}(r, \varphi)$ pour $(r, \varphi) \in \omega$ et $\overset{\circ}{y}(r, \varphi) = 0$ pour $(r, \varphi) \in \gamma$. L'opérateur de différences $\mathscr{R}$ est ici défini par la relation

$$\mathscr{R} y = \frac{1}{\rho}(\bar{r} y_{\bar{r}})_{\hat{r}} + \frac{1}{\rho^2} y_{\bar{\varphi}\varphi}, \quad (r, \varphi) \in \omega.$$

En utilisant les formules de différences de Green, on est en mesure de vérifier si les opérateurs A et R sont autoadjoints dans H et, de plus, si pour tout $y \in H$ on a les égalités

$$\begin{gathered} (Ay, y) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2-1} a_1 \overset{\circ}{y}{}_{\bar{r}}^2 h_1 h_2 + \sum_{j=1}^{N_2} \sum_{i=1}^{N_1-1} \frac{a_2}{\rho} \overset{\circ}{y}{}_{\bar{\varphi}}^2 \hbar_1 h_2, \\ (Ry, y) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2-1} \bar{r} \overset{\circ}{y}{}_{\bar{r}}^2 h_1 h_2 + \sum_{j=1}^{N_2} \sum_{i=1}^{N_1-1} \frac{1}{\rho} \overset{\circ}{y}{}_{\bar{\varphi}}^2 \hbar_1 h_2. \end{gathered}$$

De là et à partir de (48), (50) il résulte que les opérateurs A et R sont énergétiquement équivalents aux constantes $\gamma_1 = c_1$ et $\gamma_2 = c_2$:

$$\gamma_1 (Ry, y) \leqslant (Ay, y) \leqslant \gamma_2 (Ry, y), \quad \gamma_1 > 0. \tag{51}$$

Le problème de différences (49) peut être écrit sous forme d'une équation opératorielle

$$Au = f$$

dont l'opérateur A a été défini plus haut. Pour le résoudre, utilisons le schéma itératif implicite

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}}+Ay_k=f, \quad k=0,\ 1,\ \ldots, \quad y_0 \in H, \tag{52}$$

où $B = R$.

Il s'ensuit de la théorie générale des méthodes itératives exposée au chapitre VI que si les paramètres τ_{k+1} dans le schéma (52) sont choisis suivant les formules de la méthode de Tchébychev

$$\tau_k=\frac{\tau_0}{1+\rho_0\mu_k},$$

$$\mu_k \in \mathfrak{M}_n^*=\left\{-\cos\frac{(2i-1)\pi}{2n}, \quad 1\leqslant i\leqslant n\right\}, \quad k=1,\ 2,\ \ldots,\ n,$$

on aura alors pour l'erreur $z_n = y_n - u$ l'estimation

$$\| y_n - u \|_D \leqslant \varepsilon \| y_0 - u \|_D,$$

où $D = A$ ou $D = B$, $D = AB^{-1}A$, quant au nombre d'itérations, il correspond à l'estimation

$$n \geqslant n_0(\varepsilon) = \ln(0{,}5\varepsilon)/\ln\rho_1.$$

On a ici

$$\tau=\frac{2}{\gamma_1+\gamma_2}, \quad \rho_0=\frac{1-\xi}{1+\xi}, \quad \rho_1=\frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi=\frac{\gamma_1}{\gamma_2}.$$

Comme γ_1 et γ_2 ne dépendent pas des pas du maillage $\bar{\omega}$, le nombre d'itérations est proportionnel à $|\ln 0{,}5\varepsilon|$ et ne varie pas lors de la réduction du maillage.

Pour rechercher y_{k+1}, on obtient le problème de différences

$$\mathcal{R}y_{k+1} = -F, \quad (r, \varphi) \in \omega, \quad y_{k+1} = g, \quad (r, \varphi) \in \omega$$

avec le second membre connu $F = -\mathcal{R}y_k + \tau_{k+1}(\Lambda y_k + \psi)$. Notons que ce problème satisfait à toutes les conditions autorisant la recherche de la solution par l'une des méthodes directes, à savoir par la méthode de réduction totale exigeant $O(N_1N_2\log_2 N_2)$ opérations arithmétiques si $N_2 = 2^n$. Le nombre total d'opérations nécessaire à la recherche de la solution du problème de différences envisagé à la précision ε près peut donc être apprécié à la valeur $O(N_1N_2\log_2 N_2 \ln(2/\varepsilon))$.

Avec des hypothèses respectives on est en mesure de construire de façon analogue les méthodes itératives de résolution des problèmes de différences aux limites, posés aux paragraphes précédents, en coordonnées cylindriques et polaires.

ANNEXE

Construction du polynôme s'écartant le moins de zéro

1. Au § 2 du chapitre VI, lors de l'étude des schémas itératifs à deux couches, on a posé le problème: construire un polynôme de degré n prenant pour zéro la valeur 1 et dont le maximum du module sur le tronçon $[\gamma_1, \gamma_2]$ est minimal.

Résolvons ce problème. Il est plus commode de mener les études non pas sur le tronçon $[\gamma_1, \gamma_2]$, mais sur le tronçon $[-1, 1]$. A cette fin effectuons la substitution linéaire de la variable, transformant le tronçon $\gamma_1 \leqslant t \leqslant \gamma_2$ en le tronçon $-1 \leqslant x \leqslant 1$ et le point γ_1 en le point 1. Cette substitution prend la forme

$$t=\frac{1-\rho_0 x}{\tau_0}, \quad \tau_0=\frac{2}{\gamma_1+\gamma_2}, \quad \rho_0=\frac{1-\xi}{1+\xi}, \quad \xi=\frac{\gamma_1}{\gamma_2}.$$

Avec cette substitution, au point $t = 0$ correspond le point $x = 1/\rho_0 > 1$.

Le problème formulé plus haut est donc équivalent au problème suivant: parmi tous les polynômes de degré n acquérant au point $x = 1/\rho_0 > 1$ la valeur 1 rechercher celui qui s'écarte le moins de zéro sur le tronçon $[-1, 1]$.

C'est le problème classique de Tchébychev de la théorie d'approximation des fonctions dont la solution est bien connue, toutefois il sera utile de rechercher de nouveau cette solution. Il nous faut pour cela le théorème 1.

Théorème 1. *Quelles que soient sur le tronçon* $[-1, 1]$ *les fonctions continues* $g(x) > 0$ *et* $f(x)$, *il n'existe qu'un seul polynôme* $P_n(x)$ *de degré non supérieur à* n *tel que*

$$q_n = \max_{-1\leqslant x\leqslant 1} g(x)\,|f(x)-P_n(x)| = \min_{\left\{\substack{R_k(x)\\ k\leqslant n}\right\}} \max_{-1\leqslant x\leqslant 1} g(x)\,|f(x)-R_k(x)|.$$

Ce polynôme se caractérise complètement par la propriété suivante: le nombre de points successifs sur le tronçon $[-1, 1]$ en lesquels la fonction $g(x)\,(f(x) - P_n(x))$ prend avec des signes alternés la valeur q_n n'est pas inférieur à $n + 2$.

Transformons le problème posé en le rapprochant de celui figurant au théorème 1. En tenant compte de ce que le polynôme cherché prend la valeur 1 au point $x = 1/\rho_0$, représentons-le sous la forme

$$P_n(x) = 1-\left(\frac{1}{\rho_0}-x\right) R_{n-1}(x) = \frac{1-\rho_0 x}{\rho_0}\left[\frac{\rho_0}{1-\rho_0 x} - R_{n-1}(x)\right],$$

où $R_{n-1}(x)$ est un polynôme de degré non supérieur à $n - 1$.

Il en suit que notre problème se réduit à celui de la recherche du polynôme $R_{n-1}(x)$ de degré non supérieur à $n - 1$ fournissant la meilleure approximation uniforme de poids $g(x) = (1 - \rho_0 x)/\rho_0 > 0$ de la fonction $f(x) = \rho_0/(1 - \rho_0 x)$ sur le tronçon $[-1, 1]$.

C'est justement le problème figurant dans le théorème 1.

Aussi en vertu du théorème 1 existe-il au moins $n+1$ points $x_1, x_2, \ldots, x_{n+1}$ du tronçon $[-1, 1]$ en lesquels le polynôme cherché $P_n(x)$ acquiert la valeur q_n avec des signes alternés.

Montrons d'abord que le nombre de ces points doit être égal à $n+1$. En effet, pour qu'une fonction continue puisse prendre des valeurs q_n non nulles avec des signes alternés en plus de $n+1$ points successifs sur le tronçon $[-1, 1]$, elle doit s'annuler sur ce tronçon en n points au moins.

Comme le polynôme $P_n(x)$ est différent de celui identiquement nul, il ne peut s'annuler sur le tronçon $[-1, 1]$ qu'en n points au plus. Donc, le polynôme cherché $P_n(x)$ prend sur le tronçon $[-1, 1]$ la valeur q_n avec des signes alternés exactement $n+1$ fois.

Donnons la caractéristique de ces points. Si en un point intérieur du tronçon $[-1, 1]$ le polynôme $P_n(x)$ acquiert une valeur maximale, la dérivée $P'_n(x)$ s'annule alors en ce point. Mais le degré de $P'_n(x)$ est égal à $n-1$ et, par suite, la dérivée du polynôme cherché ne peut s'annuler qu'en $n-1$ points. Donc le polynôme cherché possède sur le tronçon $[-1, 1]$ $n-1$ points internes extrémaux, et partant, deux extrémums terminaux, c'est-à-dire

$$|P_n(-1)| = |P_n(1)| = q_n.$$

On a donc

$$P_n(\omega_j) = 0, \quad j = 1, 2, \ldots, n, \quad |P_n(x_j)| = q_n, \quad j = 1, 2, \ldots, n+1,$$

où ω_j sont les racines du polynôme, tandis que x_j les points extrémaux

$$-1 = x_{n+1} < \omega_n < x_n < \ldots < \omega_2 < x_2 < \omega_1 < x_1 = 1.$$

En outre, puisque $P_n(1/\rho_0) = 1$ et toutes les racines du polynôme $P_n(x)$ se trouvent sur le tronçon $[-1, 1]$, on a $P_n(1) = q_n$ et, partant, les égalités

$$P_n(x_j) = (-1)^{j-1} q_n, \quad j = 1, 2, \ldots, n+1 \tag{1}$$

se vérifient. On a le lemme 1.

L e m m e 1. Le polynôme $P_n(x)$, *qui parmi tous les polynômes de degré* n *prenant la valeur* 1 *pour* $x = 1/\rho_0$ *s'écarte le moins de zéro sur le tronçon* $[-1, 1]$, *satisfait à l'équation différentielle*

$$(1 - x^2)(P')^2 = n^2 (q_n^2 - P^2). \tag{2}$$

En effet, comme il a été montré ci-dessus, les points $x_2, x_3, \ldots, x_n$ sont des zéros simples du polynôme $P'_n(x)$. Ces points sont apparemment des zéros doubles du polynôme $q_n^2 - P_n^2(x)$, or on a montré que les points $x_{n+1} = -1$ et $x_1 = 1$ sont des zéros simples de ce polynôme. Donc, les polynômes $(1 - x^2) \times \times (P'_n(x))^2$ et $q_n^2 - P_n^2(x)$ de degré 2 possèdent les mêmes zéros. Par conséquent, ils sont proportionnels, c'est-à-dire

$$(1 - x^2)(P'_n)^2 = c\,(q_n^2 - P_n^2(x)).$$

En égalant les coefficients des puissances supérieures en x des deux polynômes, on obtient $c = n^2$. Le lemme est démontré.

2. Passons à la construction du polynôme $P_n(x)$ sur la base de l'équation (2). Cette équation, outre la fonction inconnue $P_n(x)$, comprend également le paramètre inconnu q_n. Nous ne fixerons pas séparément les conditions complémentaires déterminant de façon univoque la solution de l'équation (2) mais utiliserons toute l'information connue se rapportant à $P_n(x)$.

Etudions d'abord l'équation (2) sur le tronçon $[-1, 1]$. Dans ce cas $|P_n(x)| \leqslant q_n$, et, par conséquent, du premier et du second membre de l'équation (2) il est possible d'extraire une racine

$$\pm \frac{dP}{\sqrt{q_n^2 - P^2}} = n \frac{dx}{\sqrt{1 - x^2}}, \quad 0 \leqslant x \leqslant 1. \tag{3}$$

Etudions le premier membre de (3). Si $P_n(x_{j+1}) = q_n$, alors avec la variation de x de x_{j+1} à x_j la fonction $P_n(x)$ décroît de q_n à $-q_n$. La différentielle dP

est dans ce cas négative et, par suite, dans le premier membre de l'équation (3) il faut adopter le signe moins. De façon analogue on constate que si $P_n(x_{j+1}) = -q_n$, il faut choisir le signe plus. Compte tenu de (1), on obtient que sur le tronçon $[x_{j+1}, x_j]$ l'équation (3) doit être écrite sous la forme

$$(-1)^{j-1}\frac{dP}{\sqrt{q_n^2-P^2}} = n\frac{dx}{\sqrt{1-x^2}}, \quad x\in[x_{j+1}, x_j], \quad j=1, 2, \ldots, n. \tag{4}$$

Obtenons maintenant l'expression de $P_n(x)$ sur le tronçon $[-1, 1]$. Soit x un point quelconque du tronçon $[-1, 1]$ et, pour fixer les idées, admettons que x appartient, par exemple, au tronçon $[x_{k+1}, x_k]$.

Intégrons le second membre de l'équation (4) en x de x à 1. Il vient

$$n\int_x^1 \frac{dx}{\sqrt{1-x^2}} = n \arcsin x \Big|_x^1 = n \arccos x.$$

Intégrons le premier membre de l'équation (4). Quand x varie de x_{j+}. à x_j, la fonction $P(x)$ varie de $P(x_{j+1}) = (-1)^j q_n$ jusqu'à $P(x_j) = (-1)^{j-1}q_n$. Donc

$$(-1)^{j-1}\int_{P(x_{j+1})}^{P(x_j)} \frac{dP}{\sqrt{q_n^2-P^2}} = \int_{-q_n}^{q_n} \frac{dP}{\sqrt{q_n^2-P_n^2}} = \arcsin\frac{P}{q_n}\Big|_{-q_n}^{q_n} = \pi.$$

Ensuite, en intégrant le premier membre de (4) de $P(x)$ à $P(x_k)$, il vient

$$(-1)^{k-1}\int_{P(x)}^{P(x_k)} \frac{dP}{\sqrt{q_n^2-P_2}} = \int_{(-1)^{x-1}P(x)}^{q_n} \frac{dP}{\sqrt{q_n^2-P}} = \arccos(-1)^{k-1}\frac{P(x)}{q_n}.$$

Vu que

$$\int_x^1 \frac{dx}{\sqrt{1-x^2}} = \int_x^{x_k} \frac{dx}{\sqrt{1-x^2}} + \sum_{j=1}^{k-1}\int_{x_{j+1}}^{x_j} \frac{dx}{\sqrt{1-x^2}},$$

on obtient finalement

$$n \arccos x = (k-1)\pi + \arccos(-1)^{k-1}\frac{P(x)}{q_n}. \tag{5}$$

Il en résulte que

$$P_n(x) = q_n \cos(n \arccos x), \quad |x| \leqslant 1. \tag{6}$$

En posant dans (5) $x = \omega_k \in [x_{k+1}, x_k]$, on obtient les racines du polynôme $P_n(x)$

$$\omega_k = \cos\frac{(2k-1)\pi}{2n}, \quad k=1, 2, \ldots, n.$$

La formule (6) définit le polynôme $P_n(x)$ pour $x \in [-1, 1]$. Cherchons la forme du polynôme $P_n(x)$ pour $|x| \geqslant 1$ et déterminons q_n. A cette fin notons que

$$\omega_{n-k+1} = \cos\left(\pi - \frac{2k-1}{2n}\pi\right) = -\omega_k, \quad k=1, 2, \ldots, n.$$

Donc $P_n(-x) = (-1)^n P_n(x)$ et, par suite, il suffit de déterminer $P_n(x)$ pour $x \geqslant 1$.

Etudions l'équation (2) pour $x \geqslant 1$. Dans ce cas il faut la récrire de la façon suivante:

$$(x^2-1)(P')^2 = n^2 (P^2 - q_n^2), \quad x \geqslant 1.$$

Comme $x \geqslant 1$, $P(x) \geqslant q_n$, et la fonction croît. Donc en extrayant la racine, il vient

$$\frac{dP}{\sqrt{P^2-q_n^2}} = n \frac{dx}{\sqrt{x^2-1}}.$$

Lors de l'intégration du second membre de cette équation de 1 à x le premier membre s'intégrera de q_n à $P_n(x)$. Par conséquent,

$$\int\limits_{q_n}^{P_n(x)} \frac{dP}{\sqrt{P^2-q_n^2}} = \ln\left(\frac{P_n(x)}{q_n} + \sqrt{\frac{P_n^2(x)}{q_n^2}-1}\right) = \operatorname{arcch} \frac{P_n(x)}{q_n} =$$

$$= n \int\limits_1^x \frac{dx}{\sqrt{x^2-1}} = n \ln\left(x+\sqrt{x^2-1}\right) = n \operatorname{arcch} x. \tag{7}$$

On en déduit que

$$P_n(x) = q_n \operatorname{ch}(n \operatorname{arcch} x), \quad x \geqslant 1.$$

Comme $P_n(x) = (-1)^n P_n(-x)$, pour $x \leqslant 1$ on obtient

$$P_n(x) = (-1)^n q_n \operatorname{ch}(n \operatorname{arcch}(-x)) = q_n \operatorname{ch}(n \operatorname{arcch} x), \quad x \leqslant -1.$$

Donc, pour $|x| \geqslant 1$ le polynôme $P_n(x)$ acquiert l'expression suivante:

$$P_n(x) = q_n \operatorname{ch}(n \operatorname{arcch} x), \quad |x| \geqslant 1. \tag{8}$$

Cherchons maintenant q_n. En posant dans (8) $x = 1/\rho_0$ et compte tenu de ce que $P_n(1/\rho_0) = 1$, il vient

$$q_n = 1/\operatorname{ch}(n \operatorname{arcch}(1/\rho_0)).$$

D'autre part, en posant dans (7) $x = 1/\rho_0$, on obtient

$$\ln \frac{1+\sqrt{1-q_n^2}}{q_n} = n \ln \frac{1+\sqrt{1-\rho_0^2}}{\rho_0} = n \ln \frac{1}{\rho_1},$$

où

$$\rho_1 = \frac{\rho_0}{1+\sqrt{1-\rho_0}} = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}, \quad \rho_0 = \frac{2\rho_1}{1+\rho_1^2}.$$

Par conséquent,

$$q_n = \frac{1}{\operatorname{ch}\left(n \operatorname{arcch} \frac{1}{\rho_0}\right)} = \frac{2\rho_1^n}{1+\rho_1^{2n}} < 1. \tag{9}$$

En réunissant (6) et (8), il vient

$$P_n(x) = q_n T_n(x) = T_n(x)/T_n(1/\rho_0), \tag{10}$$

où

$$T_n(x) = \begin{cases} \cos(n \arccos x), & |x| \leqslant 1, \\ \operatorname{ch}(n \operatorname{arcch} x), & |x| \geqslant 1. \end{cases}$$

Le polynôme $T_n(x)$ est appelé polynôme de Tchébychev de première espèce de degré n.

Ainsi, le problème posé est complètement résolu. Sa solution s'obtient au moyen des formules (9), (10). En revenant à la variable t, on aboutit au polynôme cherché

$$Q_n(t) = P_n\left(\frac{1-\tau_0 t}{\rho_0}\right) = q_n T_n\left(\frac{1-\tau_0 t}{\rho_0}\right),$$

qui s'écarte le moins de zéro sur le tronçon $[\gamma_1, \gamma_2]$

BIBLIOGRAPHIE

1. *Гельфонд А. О.* Исчисление конечных разностей.— М.: Наука, 1967.
2. *Карчевский М. М., Ляшко А. Д.* Разностные схемы для нелинейных задач математической физики.— Казань: Ротапринт, изд. Каз. гос. ун., 1976.
3. *Красносельский М. А., Вайникко Г. М. и др.* Приближенное решение операторных уравнений.— М.: Наука, 1969.
4. *Марчук Г. И.*, Методы вычислительной математики.— Новосибирск: Наука, 1973. (G. Marchouk, Méthodes de calcul numérique, Editions Mir, Moscou, 1980).
5. *Оганесян Л. А., Ривкинд В. Я., Руховец Л. А.*, Вариационно-разностные методы решения эллиптических уравнений, ч. 1 и 2.— В сб.: Дифференциальные уравнения и их применение, вып. 5, Вильнюс, Пяргале, 1973, вып. 8, Вильнюс, Пяргале, 1974.
6. *Ortega J., Rheinbold W.* Iterative Solution of Nonlinear Equations in Several Variables, New York, 1970.
7. *Самарский А. А.* Введение в теорию разностных схем.— М.: Наука, 1971 (имеется библиография до 1971 г.).
8. *Самарский А. А.* Теория разностных схем.— М.: Наука, 1977.
9. *Самарский А. А., Гулин А. В.* Устойчивость разностных схем.— М.: Наука, 1973.
10. *Самарский А. А., Андреев В. Б.* Разностные методы для эллиптических уравнений.— М.: Наука, 1976 (A. Samarski, V. Andréev, Méthodes aux différences pour équations elliptiques, Editions Mir, Moscou, 1978).
11. *Самарский А. А., Карамзин Ю. Н.*, Разностные уравнения.— М.: Знание, 1978.
12. *Фаддеев Д. К., Фаддеева В. Н.* Вычислительные методы линейной алгебры.— М.: Физматгиз, 1963.
13. *Wasow W., Forsythe G.* Finite-Difference Methods for Partial Differential Equations, New-York.
14. *Young D. M.* Iterative Solution of Large Linear Systems : New-York, London: Acad. Press, 1971.

INDEX ALPHABÉTIQUE

Accélération de la convergence 383
Algorithme de la transformation discrète de Fourier 179

Coefficients de balayage 80
Correction 286

Dérivée Gâteau 235, 532
Différence centrale 26
— progressive 26
— régressive 26

Elément propre de l'opérateur 244
Equation caractéristique 50
— aux différences 30
— de mailles 30
— de mailles à coefficients constants 33
Espaces de Banach 233

Fonction de Green de l'opérateur de différences 259
— de maille 25
— de maille vectorielle 26
Formules de Green au sens de différences finies 253

Identités aux sens de différences finies 252

Maillage 23
— carré 25
— irrégulier 25
— rectangle 25
— régulier 24
Méthode du balayage 78
Méthode du balayage cyclique 92
— — en flux 89
— — matriciel 111
— — monotone 100
— — non monotone 100
— — orthogonal 121
— de réduction 131
— de séparation des variables 179
— de stationnarisation 279
— itérative des corrections conjuguées 379
— — de descente par gradient 540
— — des directions alternées 459
— — des erreurs conjuguées 379
— — de la plus grande pente 362
— — des gradients conjugués 378
— — des moindres corrections 365
— — des moindres erreurs 366, 497
— — des moindres résidus 364, 511
— — de Newton-Kantorovitch 535
— — des résidus conjugués 379
— — de Seidel 390
— — simple 305
— — de stationnarisation à trois couches 344
— — de surrelaxation 399
— — de Tchébychev 290
— — triangulaire 412
— — triangulaire alternée 420
Méthodes itératives à deux étapes 538
— — du type variationnel 353

Nœuds 323
— frontières 323
Noyau de l'opérateur 236

Opérateur adjoint 238
— autoadjoint 238
— commutatif 236
— continu 235

Opérateur défini positif 240
— de différences 26
— énergétiquement équivalent 240
— monotone 240, 529
— fortement monotone 240, 530
— rigoureusement monotone 240
— normal 238
— de passage 286
— potentiel 541
— résolvant 286

Pas du maillage 24
Points frontières 25
Polynôme de Tchébychev de première espèce 60
— — de seconde espèce 60
Première transformation de Gauss 318
Problème de valeurs propres 65
Propriété asymptotique 360
Principe de régularisation 562

Rayon numérique de l'opérateur 239, 315
— spectral 237, 401
Régularisateur 563
Réseau 23

Schéma aux différences 24
— itératif à deux couches 280
— itératif à trois couches 281
Schémas itératifs à opérateur factorisé 566
Solution généralisée 247, 507
— normale 246, 507
Spectre de l'opérateur 244
Stabilité avec information à priori 347
— sous le rapport des calculs 296
Stencil 26

Valeur propre de l'opérateur 244
Variation des constantes 44

TABLE DES MATIÈRES

Préface 4

Introduction 9

Chapitre I. MÉTHODES DIRECTES DE RÉSOLUTION DES ÉQUATIONS AUX DIFFÉRENCES 23

§ 1. Equations de mailles. Notions générales 23
§ 2. Théorie générale des équations aux différences linéaires 37
§ 3. Solution des équations linéaires à coefficients constants 49
§ 4. Equations de second ordre à coefficients constants . . . 57
§ 5. Problèmes de différences de valeurs propres 65

Chapitre II. MÉTHODE DE BALAYAGE 77

§ 1. Méthode du balayage pour les équations triponctuelles 77
§ 2. Variantes de la méthode du balayage 89
§ 3. Méthode du balayage pour les équations pentaponctuelles 104
§ 4. Méthode du balayage matriciel 111

Chapitre III. MÉTHODE DE RÉDUCTION TOTALE 131

§ 1. Problèmes aux limites pour les équations vectorielles triponctuelles 131
§ 2. Méthode de réduction totale pour le premier problème aux limites 141
§ 3. Exemples d'application de la méthode 156
§ 4. Méthode de réduction totale appliquée à d'autres problèmes aux limites 162

Chapitre IV. MÉTHODE DE SÉPARATION DES VARIABLES . . . 179

§ 1. Algorithme de la transformation discrète de Fourier 179
§ 2. Résolution de problèmes de différences par la méthode de Fourier 202
§ 3. Méthode de réduction non totale 216

Chapitre V. APPAREIL MATHÉMATIQUE DE LA THÉORIE DES MÉTHODES ITÉRATIVES 231

§ 1. Eléments d'information sur l'analyse fonctionnelle . . . 231
§ 2. Schémas aux différences considérés comme des équations opératorielles 249
§ 3. Notions générales sur la théorie des méthodes itératives 278

Chapitre VI. MÉTHODES ITÉRATIVES À DEUX COUCHES . . 287

§ 1. Position du problème sur le choix des paramètres d'itération 287
§ 2. Méthode de Tchébychev à deux couches 290
§ 3. Méthode itérative simple 305
§ 4. Cas d'opérateur non autoadjoint. Méthode itérative simple 308
§ 5. Exemples d'application des méthodes itératives 319

Chapitre VII. MÉTHODES ITÉRATIVES À TROIS COUCHES . . . 337

§ 1. Appréciation de la vitesse de convergence 337
§ 2. Méthode semi-itérative de Tchébychev 341
§ 3. Méthode de stationnarisation à trois couches 344
§ 4. Stabilité des méthodes à deux et à trois couches avec information à priori 347

Chapitre VIII. MÉTHODES ITÉRATIVES DU TYPE VARIATIONNEL 353

§ 1. Méthode du gradient à deux couches 353
§ 2. Exemples de méthodes du gradient à deux couches . . . 362
§ 3. Méthodes des directions conjuguées à trois couches . . . 369
§ 4. Exemples de méthodes à trois couches 378
§ 5. Accélération de la convergence des méthodes à deux couches au cas d'un opérateur autoadjoint 383

Chapitre IX. MÉTHODES ITÉRATIVES TRIANGULAIRES 390

§ 1. Méthode de Seidel 390
§ 2. Méthode de surrelaxation 399
§ 3. Méthodes triangulaires 412

Chapitre X. MÉTHODE TRIANGULAIRE ALTERNÉE 420

§ 1. Théorie générale de la méthode 420
§ 2. Problèmes aux limites discrets pour les équations elliptiques dans un rectangle 434
§ 3. Méthode triangulaire alternée de résolution des équations elliptiques dans un domaine arbitraire 450

Chapitre XI. MÉTHODE DES DIRECTIONS ALTERNÉES . . 459

§ 1. Méthode des directions alternées au cas de commutativité 459
§ 2. Exemples d'application de la méthode 468
§ 3. Méthode des directions alternées dans le cas général . . . 481

Chapitre XII. MÉTHODES DE RÉSOLUTION DES ÉQUATIONS À OPÉRATEURS DÉGÉNÉRÉS DE SIGNES INDÉTERMINÉS 487

§ 1. Equations à opérateur réel de signe indéterminé . . . 487
§ 2. Equations avec opérateur complexe 499
§ 3. Méthodes itératives générales pour les équations avec opérateur dégénéré 507
§ 4. Méthodes spéciales 518

Chapitre XIII. MÉTHODES ITÉRATIVES DE RÉSOLUTION DES ÉQUATIONS NON LINÉAIRES 529

§ 1. Méthodes itératives. Théorie générale 529
§ 2. Méthodes de résolution des schémas aux différences non linéaires 544

Chapitre XIV. EXEMPLES DE RÉSOLUTION DES ÉQUATIONS ELLIPTIQUES DE MAILLES 562

§ 1. Procédés de construction des schémas itératifs implicites 562
§ 2. Systèmes d'équations elliptiques 572

Chapitre XV. MÉTHODES DE RÉSOLUTION DES ÉQUATIONS ELLIPTIQUES EN COORDONNÉES CURVILIGNES ORTHOGONALES 580

§ 1. Position des problèmes aux limites pour des équations différentielles 580
§ 2. Résolution des problèmes de différences en coordonnées cylindriques 586
§ 3. Résolution des problèmes de différences dans le système de coordonnées polaires 600

Annexe 617

Bibliographie 622

Index alphabétique 623

